Abstract—Assistive speech-based technologies can improve the quality of life for people affected with dysarthria, a motor speech disorder. In this paper, we explore multiple ways to improve Gaussian mixture model (GMM) and deep neural network (DNN) based hidden Markov model (HMM) automatic speech recognition (ASR) systems for TORGO dysarthric speech database. Our work shows significant improvements over the previous attempts in building such systems in TORGO. We trained speaker-specific acoustic models by tuning various acoustic model parameters, using speaker normalized cepstral features and building complex DNN-HMM models with dropout and sequence discrimination strategies. The DNN-HMM models for severe and severe-moderate dysarthric speakers were further improved by leveraging specific information from dysarthric speech to DNN models trained on audio files from both dysarthric and normal speech, using generalized distillation framework. To the best of our knowledge, this paper presents the best recognition accuracies for TORGO database till date. Index Terms—dysarthria, multitask, distillation, acoustic model, DNN.

II. TORGO DYSARTHRIC SPEECH DATABASE TORGO is a popular dysarthric speech database of aligned acoustic and articulatory recordings from 15 speakers [20]. Eight of these speakers (5 males, 3 females) have dysarthria. The remaining seven are control speakers (4 males, 3 females) without any speech disorders [38]. The details of the dysarthric and control speakers in the database is given in Table I. For each of the eight dysarthric speakers, the severity level of the disorder was evaluated by a speech-language pathologist. This evaluation was made in terms of their overall clinical intelligibility and motor functionality of the articulators, as per Frenchay dysarthria assessment [39]. The database consists of recordings of single words, sentences, description of contents in a photograph by the speakers. Single words consisted of English digits, international radio alphabets, twenty most frequent words in British National Corpus (BNC), and a set of words selected by Kent et al. [40] to account for relevant phonetic contrasts. The sentences were taken from Yorkston-Beukelman assessment of intelligibility [41] and the TIMIT database [42]. To include dictation style speech in to the database, subjects were asked to describe the contents of a few photographs in his/her own words. Approximately three hours of speech was recorded in this manner from multiple sessions. On an average, 415 and 800 utterances were recorded from each dysarthric and control speaker respectively. The precise number of recordings from each speaker is given in Table I.

In this paper, we consider three different training setups for acoustic modeling as mentioned in [25]. They include training the acoustic model on (a) dysarthric data alone, (b) control data alone and (c) merged dysarthric and control data, as shown in Fig.1. Mengistu et al. [25] observed that combining dysarthric and control data for acoustic modeling is better when compared to the other two training strategies. Hence, we will be following the combined training strategy in this paper in order to be consistent with [25], [32]. However, the DNN-HMM acoustic model recognition performance for all the three training setups is given in section V. The Kaldi recipe1 mentioned in [32] is used for building various speaker specific acoustic models in the combined training data framework, by excluding the speech data from the test dysarthric speaker. For example, if F01 dysarthric speaker is the test subject, then as per the details given in Table I, only 5358 audio files from the remaining seven dysarthric speakers and 10931 audio files from all control speakers will be used for training the acoustic model. In the following subsections, details of various proposed modifications to GMMHMM acoustic modeling are discussed. These modifications are applicable irrespective of the training setups used.

1. Frame Rate

Using frame length of 25ms, 13-dimensional Mel frequency cepstral coefficients (MFCC) are extracted. Mengistu et al. [25] used frame shift of 15ms for severely dysarthric speakers and 10ms for moderate-mild dysarthric speakers and all control speakers. España-Bonet et al. [32] used frame shift of 15ms for all dysarthric speakers and 10ms for control speakers. The increase in the frame shift suggested in [25], [32] was meant to compensate for the slow speaking rate of dysarthric speakers and homogenize the difference between dysarthric and control speech. We verify this claim in Table II. The MFCC features for dysarthric speech was extracted in two ways, (a) with 10ms frame shift and (b) with 15ms frame shift. Using these features, monophone and triphone GMM-HMMs are trained in combined framework for the case when F01 dysarthric speaker is the test subject. A 3-state HMM model was trained for each phoneme and a total of 1000 Gaussians were used. We used 1800 context dependent states and 9000 Gaussians for triphone modeling as mentioned in [32]. From Table II, we infer that while increasing in frame shift to 15ms for dysarthric speech is beneficial as claimed in [25], [32].

B. Word Position Dependency of Phones

The list of phones used in acoustic modeling can be expanded by taking into account the position of phones within a word. For instance, for a phone aa, this extended phone list will include aa\_B, aa\_E, aa\_I and aa\_S, indicating whether the phone occurs in the beginning, end, internal of a word or as a singleton phone respectively. This word position dependent phone set was used in [32] for acoustic modeling. This method is generally adopted in Kaldi toolkit [43] to account for coarticulation in speech. Dysarthric speech is characterized by slow speaking rate, slurring, mumbling, elongation of words and stop-gaps. The dysarthric speakers may not always be able to complete the word pronunciations. These factors lead to the phones occurring in isolation than as a part of a word as in the case of normal speech. Thus with very low co-articulation levels in dysarthric speech, we infer that using word position dependent phones in acoustic modeling is an overkill. Table II verifies our claim and shows that using word position independent phones give relative WER reduction of 35.88% when compared to [32]. Hence, in our experiments, we have used frame shift of 15ms for dysarthric speakers and word position independent phones for acoustic modeling. The WER for monophone GMM-HMM models for all test dysarthric speakers using the proposed frame shift and phone set is given in Table III.

C. Number of Context Dependent States

As mentioned earlier, speaker dependent triphone models are trained on audio files from all control speakers and dysarthric speakers, excluding the utterances from test dysarthric speaker. Hence, each dysarthric speaker’s GMMHMM will be dependent on the train data characteristics like number of train utterances, speakers involved etc. This necessitates the tuning of the number of context dependent states (tied states). Also, these context dependent states are further carried on for neural network modeling. Hence obtaining a properly tuned GMM-HMM model is of paramount importance. In [32], the number of tied states was fixed at 1800 for all dysarthric speaker acoustic models. The number of tied states was varied form 400 to 1800 for each speaker dependent triphone model in Table IV. The total number of Gaussians was fixed at 9000 as mentioned in [32]. From Table IV, we make the following observations: • Tuning the number of tied states has significantly decreased the WER of triphone models, compared to [32]. • Using lower number of tied states gives better performance. Dysarthric speech has low degree of coarticulation with phones often occurring in isolation. Tied states are meant to capture co-articulation effects of speech. Hence, dysarthric speakers requires only fewer tied states for acoustic modeling.

D. Dimensionality of LDA Features

The first and second order derivatives of 13-dimensional MFCC are augmented to itself to generate 39-dimensional feature vectors, which is then mean normalized at speaker level. Nine consecutive frames of this feature is then spliced together, projected to a M-dimensional vector using linear discriminant analysis (LDA) and further diagonalized by maximum likelihood linear transformation (MLLT). The Mdimensional features are called LDA features. M was fixed at 40 in [32]. Each dimensional of the LDA feature represents the principal directions of variations of the feature vector.

In Table V, we varied M from 25 to 40 in steps of 5. The WER of speaker dependent GMM-HMMs trained on Mdimensional LDA features are reported. A reduction in the dimensionality of LDA feature shows improved WER in all cases. In our experiments, we have chosen the dimensionality of LDA features as 30. E. Speaker Normalized FMLLR Features Methods to reduce acoustic variabilities due to speaker influence are collectively referred to as speaker normalization methods. Speaker normalizing transforms like feature-space maximum likelihood linear regression (FMLLR) can be applied on LDA features to remove the speaker variabilities in the data. The FMLLR features can then be used for compact acoustic modeling. Since we chose 30-dimensional LDA features in our experiments, the FMLLR features will also be 30-dimensional. In [32], 40-dimensional FMLLR features were used. Table VI reports the WER of GMM-HMMs trained on MFCC, LDA and FMLLR features. It can be seen that removal of speaker variabilities using FMLLR reduces the WER. Compared to [32], the proposed reduced dimensional FMLLR transform gave 24.64% reduction in WER across all dysarthric test speakers.

IV. DNN-HMM ACOUSTIC MODELING

In dysarthric ASR, studies show that neural networks handles complex acoustic modeling, extraction of discriminative features, pronunciation error correction and transferring knowledge from normal speech corpus. In this section, we explore various techniques to improve upon the DNN-HMM results reported in [32]. We used 30-dimensional FMLLR features for training and decoding DNN models. The number of units in the output softmax layer was equal to the number of tied states in phonetic decision tree. The frame-level alignment information was obtained from the GMM-HMM model trained on FMLLR features. Initially, the entire training data was randomized at frame level. The DNN parameters were initialized by layer-wise restricted Boltzmann machine (RBM) pretraining. Supervised training uses stochastic gradient descent with a mini-batch size of 256 frames and learning rate of 0.008. The FMLLR features were stacked over a context window of 11 frames (±5) and are fed as input to the DNN. Weighted finite state transducer (WFST) based graph generated for GMM-HMM model was used for decoding the DNN models by scaling DNN posteriors with class priors computed from alignments. A. Number of Hidden Layers and Nodes España-Bonet et al. [32] reported recognition results on DNN models trained with 6 hidden layers and 1024 neurons per layer. We reduced the complexity of the DNN by reducing the number of hidden layers and nodes as shown in Table VII. It can be seen that a less complex model gives 27.72% relative reduction in WER compared to [32]. We employed DNNs with 4 hidden layers with sigmoid activation functions and 1024 neurons in each layer in our experiments. The recognition results of various speaker-specific DNN-HMM models are given in Tables VIII and IX.

B. Dropout for DNN Generalization

The DNN model must be able to generalize the information learned from other dysarthric speakers to avoid overfitting to the train data. Applying dropout has been shown to control overfitting and increase the generalizing ability of the DNNs [44], [45]. Dropout technique randomly omits a certain percentage (α) of the neurons in each hidden layer for each presentation of samples during DNN training. Hence during training, a random combination of (1 − α) hidden neurons has to perform well in the absence of omitted neurons. This reduces the dependency of each neuron on other neurons to detect patterns. In our experiments, we applied a dropout factor of 0.2 for the first four DNN training epochs. This has reduced the average relative WER by 14.12% as shown in Table VIII.

C. Sequence-Discriminative Training

The DNNs models seen so far were trained to classify individual frames based on the cross-entropy criterion, which minimizes the expected frame error. But ASR is a sequence classification problem. Sequence-discriminative training [46] tries to match the maximum a posteriori (MAP) decision rule of continuous speech recognition by considering sequence or inter-frame constraints from HMMs, dictionary and language model. Better recognition accuracy can be achieved if DNNHMMs are trained using sequence-discriminative criteria like maximum mutual information (MMI) [47], boosted MMI (BMMI) [48], minimum phone error (MPE) [49] or minimum Bayes risk (MBR) [50]. State MBR (sMBR) minimizes the expected state error by taking into consideration both HMM topology and language model. In our experiments, we applied 6 iterations of sMBR on top of a DNN trained using crossentropy criterion. In [32], improvements were reported when sequence-discriminative training was applied. But dropout was not applied to these DNNs.

In this paper, we report results when sequencediscriminative training is applied on top of DNN trained with and without dropout. It can be seen from Table IX that sequence-discriminative training improves the recognition accuracy of DNNs, with or without the application of dropout. The proposed DNN models, when compared to the sMBR applied DNN in [32], shows 17.62% relative reduction in WER, across all the dysarthric test speakers. Specifically, for severe dysarthric speakers, a relative reduction of 20.11% in WER was observed when compared to [32].

V. IMPROVING SEVERE AND SEVERE-MODERATE DYSARTHRIC SPEAKER DNN-HMM MODELS In this section, we propose a generalized distillation framework [37] to further improve the DNN-HMM acoustic model performance of severe and severe-moderate dysarthric speakers reported in Table VIII. The motivation and relevance for the proposed approach is explained in the following subsections. A. Motivation In section III, we mentioned about the three different training strategies possible in TORGO database. Table X shows the WER of DNN-HMM models trained using these strategies, viz., control data alone, dysarthric data alone and combining dysarthric and control data. The DNN models in the first two cases has 2 hidden layers and 512 neurons in each layer. The combined DNN model has 4 hidden layers with 1024 neurons per layer. All DNN models are trained with dropout factor of 0.2 in the first four iterations. Note that the “combined” DNN

model in Table X is the same as the DNN model with dropout applied in Table VIII. The following observations can be made from Table X:

* Severe dysarthric speakers (F01, M01, M02, M04) shows better recognition accuracy when DNNs are trained with dysarthric speech data alone. An exception to this observation is M04 dysarthric speaker.
* Severe-moderate dysarthric speaker (M05) also gives improved recognition accuracy for DNNs trained only on dysarthric speech data.
* For severe and severe-moderate dysarthric speakers, DNNs trained on pooling data from dysarthric and control speech is not benefiting from the additionally supplied control speaker data. As the severity level of dysarthria increases, the differences between dysarthric and control speech increases. Hence the specificities of the dysarthric speech may not be captured by DNNs trained on combined speech data.
* For moderate (F03) and mild (F04, M03) dysarthric speakers, DNNs trained only on control speech data performs better than those trained on dysarthric speech data alone. This indicates that the subjects with moderate and mild dysarthria are closer to control speakers and produce almost un-impaired speech as reported in [25].
* Due to the above observation, pooling data from dysarthric and control speakers for DNN training improves the recognition performance of moderate and mild dysarthric speakers.

The data insufficiency problem in building dysarthric ASR is normally tackled by pooling together data from control speech of the same or any other standard database with the dysarthric speech [5], [11], [21]–[27]. But in the case of TORGO, for severe and severe-moderate dysarthric speakers, this approach seems to degrade the recognition accuracy when compared to models built on dysarthric speech alone. Also, most of these data pooling strategies have to be monitored manually as the amount of control speech added must not bias the resultant combined DNN model towards the control speech. Hence, we need a framework which can:

* Provide an intelligent technique to transfer knowledge about dysarthric speech to control speech.
* Train DNN models on pooled dysarthric and control speech without biasing the model towards control speech.

We propose a generalized distillation framework which transfers knowledge about dysarthric speech to the combined DNN model to resolve this issue. The proposed method attunes the combined DNN model towards the dysarthric speech, while still benefiting from the inclusion of control speech data

for building a complex DNN model. The details about the proposed framework is given in the following subsections.

B. Relation to Prior Work

Generalized distillation is a framework where a machine teaches another machine about a task. The intelligent teacher machine passes the information about the task to be learned to the student machine [51], [52]. This method of distilling knowledge was applied in neural networks in [53]. Generalized distillation framework was used for noise robust speech recognition in [54]. Here, a student DNN trained on noisy features exploits the information about corresponding clean speech from a teacher DNN trained on clean speech. In [55], we extended this concept to speaker normalization, where student DNN trained on un-normalized MFCC features uses the information about the corresponding FMLLR features from the teacher DNN trained on FMLLR features.

C. Teacher and Student DNN Training

The proposed teacher-student framework is shown in Fig.2. A DNN model trained on dysarthric speech data alone is the teacher DNN. The student DNN is trained on both dysarthric and control speech data. Here, the task to be learned by the student DNN is the specificities of dysarthric speech. This information needs to be passed on intelligently to the student DNN by the teacher DNN. This ensures that the student DNN does not get biased to the control speech data. Once teacher DNN model is successfully trained, combined train features are passed through it to obtain posterior probabilities. These soft targets are then given to student DNN as additional information about competing classes (context dependent states) in the student DNN output layer as shown in Fig.2. The teacher DNN transfers it’s knowledge about the dysarthric speech to the student DNN in this manner. This in turn helps the student DNN to give more weightage to the dysarthric speech information.

The student DNN also receives information about context dependent states from the GMM-HMM model trained on combined data. These are called hard targets and they are used to modify the soft targets during student DNN training. The student DNN optimization is the weighted average of two objective functions, viz., cross-entropy with hard targets and cross-entropy with soft targets.

Algorithm 1 gives the details of student DNN training. The logit units of teacher and student DNN output layers are scaled by a temperature factor, T. It controls the peakiness of the softmax outputs in teacher model. An increase in T will make the targets softer. The extent of imitation by student is controlled by an imitation factor, w. The optimal value of T = 1 was found empirically. Higher values of T reduced the recognition accuracy drastically. Various values of w = {0, 0.25, 0.5, 0.75} were applied in each case and the optimum value was found empirically. We observed that increasing the imitation factor beyond 0.75 degraded the student DNN performance significantly. Giving more weightage to soft targets, reduces the influence of hard targets. The student DNN optimization will then be solely on the basis of posterior probabilities of teacher DNN. This confuses the student DNN in choosing proper context dependent state and the model collapses as a result. Combined train features are passed through the teacher DNN, with T values applied on the softmax function, to generate soft targets for the student DNN. These posterior probabilities are then pruned by a threshold value and top 50 candidates are retained. Pruning reduces the number of memory reads and overall training time. Hard targets for the student DNN are generated from frame-level alignments of GMM-HMM model trained on combined features. During student DNN training, the hard and soft targets are appended together and multitask optimization [56] is performed. Once the student DNN is optimized, the test dysarthric subject’s data is passed to it for decoding.

D. Results and Discussion

The teacher DNN (dysarthric DNN in Table X) has 2 hidden layers and 512 neurons in each layer. The distilled student DNN has 4 hidden layers with 1024 neurons per layer. Both DNN models are trained with dropout factor of 0.2 in the first four iterations. Fig.3 shows the effect of varying the value of imitation factor w, with temperature T fixed at 1, across different degrees of dysarthric speech disorder. The following observations can be made from Fig.3:

* For severe dysarthric speakers in Fig.3a, the optimum value of imitation factor is w = 0.5 for the distilled student DNN. An exception is M04 dysarthric speaker, whose optimum value is w = 0.25. In this case, the dysarthric teacher DNN, although supervises the student DNN, does not overtly allow the student DNN to imitate it. This claim is supported by the observation that an increase in the value of w beyond 0.5 degrades the recognition performance of the distilled student DNN.
* For severe-moderate dysarthric speaker in Fig.3b, the optimum value of imitation factor is w = 0.75. The distilled student DNN imitates the dysarthric teacher DNN to a large degree here.
* For moderate and mild dysarthric speakers in Fig.3c, the optimum value of imitation factor is w = 0. These speakers are closer to control speakers and produce almost un-impaired speech. Hence a dysarthric teacher DNN does not have any supervision to offer the distilled student DNN in this case. This can be further verified by observing that increasing the imitation factor of the distilled student DNN degrades its recognition performance.

Table XI compares the WER for DNN models trained on dysarthric data alone, combined data and the distilled student model trained on combined data. The results are reported for optimum value of w and T = 1 shown in Fig.3. We make the following observations from Table XI:

* For severe dysarthric speakers, the proposed distilled student DNN trained on pooled dysarthric and control speech data gave significant improvement in recognition performance compared to DNNs trained either on dysarthric speech data alone or those trained from pooled dysarthric and control speech data.
* For severe-moderate dysarthric speaker, significant performance gain was observed for the distilled student DNN when compared to the other DNN counterparts.
* For moderate and mild dysarthric speakers, the distilled student DNN did not give any performance gains.
* The proposed distillation framework improves the DNNHMM recognition performance for dysarthric speakers with higher degree of disorder.
* All dysarthric speakers showed reduction in WER either with regular DNN or distilled student DNN models trained on pooled dysarthric and control speech data

VI. CONCLUSION

In this paper, we have explored multiple ways to improve the recognition accuracies of GMM-HMM and DNN-HMM acoustic models for TORGO dysarthric speech database. For GMM-HMM, we tuned parameters like frame rate, word position dependency of phonemes, number of contextdependent states, dimensionality of LDA features and speakernormalization via FMLLR transform. Parameters like number of hidden layers, neurons per layer were varied and strategies like dropout and sequence-discrimination training were adopted for DNN-HMM training. We proposed a generalized distillation framework to further improve DNN-HMM of severe and severe-moderate dysarthric speakers by transferring the knowledge about dysarthric speech to a DNN trained on both dysarthric and control speech. Our paper shows significant improvements over previous attempts at building ASR systems for TORGO.