A Weighted Speaker-Specific Confusion Transducer-Based Augmentative and Alternative Speech Communication Aid for Dysarthric Speakers

T. A. Mariya Celin, G. Anushiya Rachel, T. Nagarajan, *Member, IEEE*, and P. Vijayalakshmi, *Senior Member, IEEE*

***Abstract*—An augmentative and alternative speech communication (AASC) aid comprises a speech recognition system and a speech synthesis system. The main challenge in developing such an aid for dysarthric speakers lies in handling errors in the text derived from the recognition system. These errors (substitution, deletion, and insertion) may be due to inability of a dysarthric speaker to utter certain phones (articulatory error) or due to inaccuracy of the models trained (modeling error). Most existing AASC approaches only focus on the articulatory errors and the ones that do address both errors, and do not differentiate between them. However, this paper performs a three-level cascaded analysis to identify and distinguish between these errors, as differentiating these errors will aid in appropriately handling them. Furthermore, analyses in the paper are independent of the syntax of utterances. Based on these analyses, weighted phone confusion transducers are formulated and used to correct erroneous text from the recognition system. The corrected text is finally synthesized by a text-to-speech synthesis system. The proposed AASC is observed to significantly reduce a word error rate of severe dysarthric speakers from 100% to 41.52%, moderate from 61.85% to 18.08%, and mild from 12.23% to 8.55%.**

***Index Terms*—AASC, dysarthria, pronunciation errors, acoustic modeling errors, confusion transducer.**

I. Introduction

DYSARTHRIA is a neuro-motor disorder that could be caused due to cerebral palsy, a traumatic brain injury, stroke, or any cerebellar disease. This causes the speech motor control system to become weak, causing slow or uncoordinated articulatory movement, resulting in unintelligible speech.

Many treatments have been proposed to improve the intelligibility of dysarthric speech depending on type of dysarthria and its severity. Treatments typically include articulation and phonological treatments, where the speakers are taught proper production or pronunciation of speech sounds and phonological rules of a language. However, these treatments are not equally effective on all dysarthric speakers and the outcome is subjective. Therefore, an artificial remedy, namely augmentative and alternative communication (AAC), through technological assistance, might be more effective and provide a better way for communication.

AAC can be accomplished using speech-output communication aids (SOCA) and speech-input speech-output communication aids (SISOCA). Speech-output communication aid requires any user to type the text that he or she wishes to speak or select an icon to convey desired information. A text-to-speech (TTS) synthesis system may then be used to convert the information (in textual form) to speech [1]. However, as reported in [2], dysarthric speakers with motor disorders prefer to speak, since it is the most natural mode of communication. Also, dysarthric speakers ought to practice speaking to improve their speaking skills, while Speech-output communication aid restricts dysarthric speakers from doing this, which could possibly lead to further deterioration in their speaking skills.

Therefore, a more viable means of AAC would be through speech-input speech-output communication aids (SISOCA). This is an augmentative and alternative speech communication (AASC) aid that takes dysarthric speech as input, and results in normal or clear speech as output, with reduced keyboard intervention. One such system is described in [3]. It recognizes dysarthric speech as isolated words, frames a message from the recognized words, and uses a TTS system to synthesize the framed message. An important factor in successful functioning of a SISOCA is the performance of the recognition system, since an erroneous recognition outcome would result in synthesis and delivery of an incorrect message. In this regard, a lot of effort is being directed globally to dysarthric speech recognition (DSR), specifically in identifying most suitable features to be used, modeling techniques, handling sparse data, and handling recognition errors. These are discussed below.

Literature reveals that the choice of acoustic features plays an important role in DSR. In this regard, [4] uses Mel

Manuscript received September 1, 2018; revised December 7, 2018; accepted December 8, 2018. Date of publication December 17, 2018; date of current version February 8, 2019. This work was supported by the Department of Science and Technology, Government of India, through the Project on Speech-Input Speech-Output Communication Aid for speakers with Cerebral palsy under Grant SEED/TIDE/027/2016/G. *(Corresponding author: T. A. Mariya Celin.)*

The authors are with Speech Lab, SSN College of Engineering, Chennai 603 110, India (e-mail: mariyacelinta@ssn.edu.in; anushiyarachelg@ssn.edu.in; nagarajant@ssn.edu.in; vijayalakshmip@ssn.edu.in).

Digital Object Identifier 10.1109/TNSRE.2018.2887089

frequency cepstral coefficients, while [5] uses articulatory features, obtained using electromagnetic articulograph (EMA), since they may be more suitable for dysarthric speech recognition. However, articulatory features are tedious and expensive to obtain.

Different modeling techniques have also been adopted for the DSR, namely hidden Markov modeling (HMM) [6], support vector machines (SVMs) [7], Kullback-Leibler divergence-based HMM [8], and deep learning methods like artificial neural network (ANN) [9] and hybrid ANN-HMM [4]. However, all these have been used for isolated word dysarthric speech recognition. Further, deep learning methods also require a large amount of training data [9], which is not always easy to obtain from dysarthric speakers. To handle the scarcity in dysarthric speech data, a speaker adaptation technique is typically used, as in [10]. However, for speakers with severe dysarthria, speaker-dependent (SD) systems would be more suitable, since the dysarthric speaker's speech characteristics would be far too different from those of unimpaired speaker's speech data that are used to train initial speaker-independent models.

Despite careful selection of appropriate features and modeling techniques, more often than not, the recognized text is prone to errors, primarily owing to pronunciation or articulatory errors of the dysarthric speaker and also due to errors introduced by the recognition system due to inaccurate training of models (recognition or modeling errors). In this regard, the errors in the DSR system ought to be corrected and this is typically done through use of dictionaries and transducers, as elaborated below.

In [11], pronunciation errors or deviations of each dysarthric speaker are identified through a perceptual analysis and a speaker-specific pronunciation lexicon is created based on this analysis. This lexicon consists of multiple pronunciations for each word, that reflect errors specific to each dysarthric speaker. Although this method handles the pronunciation errors, it does not take measures to reduce the errors introduced by the recognition system.

A combination of feature representation and feature prediction is used in [12] to correct errors of the DSR system. As a first step, speech is converted into a phone sequence using an ASR system and the recognized phone sequence is compared with a canonical sequence using a weighted finite state transducer (WFST). In the next step, a sparse linear model incorporated with the phonological knowledge from the first step is used for error prediction. However, the experiments are conducted on a limited set of isolated words and this method would not be able to handle words that are not used in the analysis.

In [13], errors in the dysarthric speech are classified as articulatory errors and recognition errors through a likelihood analysis, using product of Gaussians technique. However, here, again, when developing a speaker-specific pronunciation dictionary, only the articulatory errors are considered, leaving behind the recognition system errors.

In [14], a speaker adaptation-based continuous dysarthric speech recognition system is trained for dysarthric speakers in the Nemours database [15]. Errors that could occur for each dysarthric speaker are analyzed using a recognition system trained on unimpaired speakers' data, and corrected using metamodels and WFSTs. Although this technique shows promising results, the errors that are modeled using WFSTs are obtained by comparing the actual transcription of the sentence with its recognized transcription from the ASR system. However, if the ASR models are not accurate, the error patterns observed may not be consistent for all examples of the same sentence. Therefore, the confusion transducer may not be accurate. Further, this method can be efficiently used only for utterances that are framed with the same words and syntax as the training data (as in Nemours [15]). However, if this technique is used for a large vocabulary system, for all possible syntaxes, it will result in a huge confusion transducer, with several possible outcomes for a single phone, specific to each dysarthric speaker. This could in turn result in an increase in word error rate.

Our current research work aims to address the issues in existing methods by studying phone characteristics (error pattern) of each dysarthric speaker, in an unrestricted syntax of sentences, and test them with a different set of utterances. The intuition behind this work is that when articulation errors are consistent they are predictable [14], [16]. These predictable errors that are consistent for a specific speaker are more likely to improve the performance of the DSR system. Initially, phone error characteristics of each dysarthric speaker is studied through a three-level cascaded analysis. As an outcome of this analysis, the pronunciation errors are distinguished from the modeling errors, unlike in [14]. Distinguishing between the errors can aid in providing proper speech therapy to dysarthric speakers, based on their articulatory errors. Further, if the accuracy of the recognition system is improved, then only the difference in the modeling errors would have to be handled, while the articulatory errors would remain unaltered. Both these errors are modeled in the current work, using a phone confusion transducer, to correct erroneous text from the DSR system. Further, since the phone characteristics are studied in an unrestricted syntax of sentences, in future, if the vocabulary size of the AASC has to be increased, the phone confusion transducer can be left untouched and only the reference transcriptions for the new words would have to be added.

The rest of the paper is organized as follows: Section II discusses speech corpora used in the current work. Section III describes a three-level cascaded error analysis. Section IV provides a detailed account of the components of the AASC aid and the performance of the aid at each stage. Finally, Section V concludes the paper.

II. Speech Corpora

The experiments in the current work are validated using two dysarthric speech corpora, namely Tamil dysarthric speech corpus developed by Mariya Celin *et al.* [17] and Nemours database [15]. The Nemours database includes dysarthric speech data from 10 dysarthric speakers having varying speech intelligibility who have uttered 74 sentences each. In order to develop a speaker adaptive system for the Nemours dysarthric speakers, TIMIT speech corpus [18] is also used.

The TIMIT speech corpus contains speech data recorded from 630 speakers of eight major dialects of American English. Each speaker has recorded 10 phonetically rich sentences. The Nemours and the Tamil speech corpus contain 38 phones each (23 are common to both languages and the rest are unique to each language) and the phones are represented by the IPA notation in the current work. A brief description of the Tamil dysarthric speech corpus is given below:

A. Tamil Dysarthric Speech Corpus

The Tamil dysarthric speech corpus contains speech data in Tamil, from 22 dysarthric speakers (17 male and 5 female) having cerebral palsy with spastic quadriplegia or diplegia. Of these 22 dysarthric speakers, 12 belong to the age group of 19 to 37 years, 5 between 15 and 18 years and, 5 between 12 and 14 years. The corpus also includes 10 unimpaired speakers (5 female and 5 male), of different age groups, between 12 and 30 years. These speakers have recorded 365 utterances each, consisting of 103 words and 262 sentences (containing 2 to 6 words). These utterances are formulated such that there are sufficient examples for all the phones. The words are chosen such that the effect of the phone articulatory errors can be observed at the beginning, middle, and end of a word. These sentences do not have a fixed structure or syntax and contain a combination of common and uncommon Tamil phrases. Tamil has a total of 40 phones, of which the corpus includes phrases formulated using 38 phones, as the remaining two phones (/f/ and /au/) occur very rarely in the language.

The dysarthric speech data was collected in collaboration with the National Institute for Empowerment of Persons with Multiple Disabilities (NIEPMD), prior to which signed consents from the parents of the dysarthric speakers were obtained. The recordings are performed in two sessions, in a laboratory environment, at a sampling rate of 16 kHz. Dysarthric speakers, who could not produce connected speech with ease, uttered sentences with a maximum of up to 6 words in sequences of 2 to 3 words, some words of a sentence were uttered in isolation. Severe dysarthric speakers uttered all words in each sentence in isolation. The corpus includes time-aligned word and phonetic transcriptions. These phonetic transcriptions are initially derived using forced Viterbi alignment procedure, as described in [19], and then manually corrected. For severe speakers, phone-level segmentation is performed based on intelligible consonants in the utterance. Of the 22 dysarthric speakers from whom the data was collected, experiments are performed on 20 dysarthric speakers, as one of the mild dysarthric speakers left after session 1, and marking the time-aligned phonetic transcription for one of the severe dysarthric speakers was difficult, due to the presence of only vowel sound units. Of these 20 dysarthric speakers, 7 are classified as mild, 10 as moderate, and 3 as severe dysarthric speakers. Classification in terms of degree of dysarthria is based on the speech intelligibility scores obtained from a speech therapist at NIEPMD and a speech intelligibility assessment test [17], conducted using listeners with phonetic expertise, in a laboratory environment.

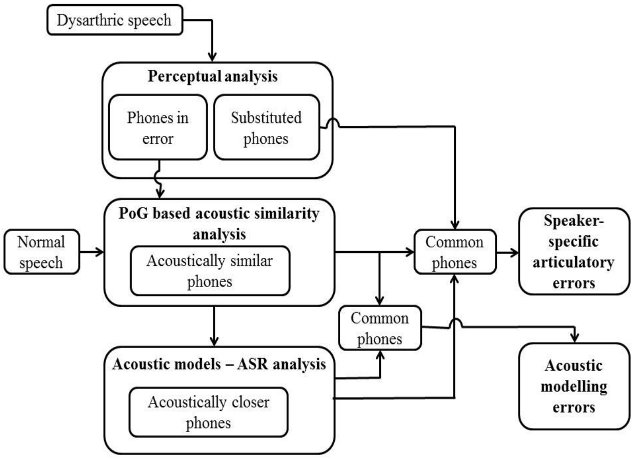
Dysarthric speech
Perceptual analysis
Phones in error
Substituted phones
Normal speech
PoG based acoustic similarity analysis
Acoustically similar phones
Common phones
Speaker-specific articulatory errors
Common phones
Acoustic models – ASR analysis
Acoustically closer phones
Acoustic modelling errors

Fig. 1. Block diagram for identifying speaker-specific articulation errors.

III. Articulatory and Modeling Error Analysis

As discussed in Section I, any DSR system is prone to, (a) pronunciation or articulatory errors and (b) recognition system or acoustic modeling errors. Therefore, the following section attempts to identify the errors specific to each dysarthric speaker and to distinguish between the consistent and predictable articulatory errors and modeling errors. To accomplish this, a three-level cascaded analysis is performed, that involves (i) a perceptual analysis, to identify the phones that are in error, (ii) a product of formant Gaussian-based acoustic similarity analysis, to validate the perceptual analysis, and finally, (iii) an ASR-based acoustic modeling analysis, to distinguish between these two errors.

Block diagram in Fig. 1 shows the complete process involved in identifying and distinguishing the articulatory errors from the acoustic modeling errors. Initially, a perceptual analysis is performed on dysarthric speech data to identify the phones that are in error and their corresponding perceptually similar phones. The outcome from this analysis is used in product of formant Gaussians-based acoustic similarity analysis, to quantify subjective analysis. A list of acoustically similar phones is obtained from this level. Finally, in the third level, an acoustic modeling analysis is performed. The errors that are consistent across all three analyses are considered to be the articulatory errors and the remaining erroneous phones, common to the product of formant Gaussian-based analysis and the ASR-based analysis, are considered to be the acoustic modeling errors. For this analysis, 200 sentences from the Tamil dysarthric corpus and 54 sentences from the Nemours database are used for each dysarthric speaker.

A. Perceptual Analysis

Perceptual analysis is performed on the Nemours and the Tamil dysarthric speech corpora, containing a total of 30 dysarthric speakers, with 30 expert listeners trained in phonetics, in a laboratory environment. Each listener is given a subset of dysarthric speakers to analyze and each dysarthric speaker is analyzed by three unique expert listeners. The dysarthric speech utterances are played back as many times as required to note down the phones that are in error and the corresponding type of error (substitution, deletion, or insertion), for each dysarthric speaker. Also, spectrograms of the erroneous phones in each utterance are compared with the corresponding spectrogram of the phone uttered by an unimpaired speaker. Errors that are common across the all 3 expert listeners alone are considered for further analysis. The agreement between the listeners is measured using Fleiss kappa measurement and the kappa value is found to be 1, that correspond to perfect agreement between listeners. Following are different error patterns observed from this perceptual analysis:

- **Substitution errors:** Phones that are substituted by each dysarthric speaker are noted separately along with their corresponding substitutions (refer Table I). These errors are grouped into three categories as discussed below:
  (i) **Fixed substitutions:** In this category, certain phones are always replaced by a set of specific phones, irrespective of the context in which they occur. This characteristic is observed in a few severe dysarthric speakers, as they tend to substitute certain phones with phones that they are able to articulate with ease.
  (ii) **Context-dependent substitutions:** Phones of this category are substituted by different phones in different contexts and in some contexts they also retain their own identity. The most common type of substitution that occurs in both corpora is substitution of a consonant in a CV (consonant-vowel) cluster by its adjacent vowel unit.
  (iii) **Random substitutions:** In most occurrences, phones of this category retain their identity. However, some examples of these phones are substituted by different ones, depending on convenience of the speaker and influence from the previously uttered context, and are hence considered to be random substitutions.
- **Insertion and deletion errors:** Fricatives, stops, and a few nasal sounds are observed to be inserted in both the corpora. However, these insertions do not occur frequently. Word-end deletions for sounds /t/ and /l/ are the most common deletion errors, however, they are still random and infrequent.

From the observations noted by the experts, the number of phones deleted, inserted, and substituted are consolidated. It is inferred that insertion, deletion, and substitution errors contribute to 19%, 12%, and 69% of the total number of errors, respectively, in the Tamil dysarthric speech corpus and 4%, 16% and, 80% of the total number of errors in the Nemours dysarthric corpus. The percentages across the categories of substitution errors varies depending on the severity of the dysarthric speaker.

Following are the overall observations made from the perceptual analysis: (i) substitution errors are the most dominant form of errors, (ii) consonant sound units are subject to errors more often than vowels, which might be due to the fact that consonants require more complex articulatory movements than vowels, (iii) consonants are substituted by both the vowels and consonants, whereas the vowels are substituted by other vowels (refer Table I), (v) though deletion and insertion errors can be perceptually observed, they are difficult to quantify as discussed in [20].

Since substitution errors are the most dominant form of errors for all the 30 dysarthric speakers, analysis based on substitution errors will be discussed further.

In this first level of analysis, a list of speech sounds exhibiting substitution error and their corresponding substituted phones are noted for all 30 dysarthric speakers. To validate the results of the perceptual analysis, a second level quantitative analysis is performed as discussed below.

B. Product of Formant Gaussian Analysis

The product of Gaussians (PoG) was initially proposed in [21] for bias estimation in a classifier and later used in [22] and [13] for acoustic similarity analyses. Since this approach showed promising results in identifying acoustically similar phones, it is used in the current work as well, to quantify the results of the perceptual analysis based on acoustic similarity. However, while previous methods [22], [13] operated in the likelihood space, the current work operates in the feature space. This is beneficial since the accuracy of the analysis is independent of the accuracy of the phone models.

For the PoG analysis, formant frequencies (F1, F2, and F3) that uniquely define each phone are initially derived using linear predictive analysis. Using these formant frequencies, formant Gaussian distributions are computed. The acoustic similarity or dissimilarity between the erroneous phones of dysarthric speaker and the phones of unimpaired speaker is derived based on the amount of overlap between their formant Gaussians. Consider a phone, /θ/ (unvoiced fricative), belonging to the context-dependent substitution category for a severe dysarthric speaker in the Tamil dysarthric speech corpus. Following are the steps involved in identifying the phones that are acoustically similar to this erroneous phone:

Step 1 (Choosing the Unimpaired Speaker): To find the closest unimpaired speaker, apart from age, gender, and dialect, formant distributions (considering all the examples and all frames in each example) between the phones of a dysarthric speaker and corresponding phones uttered by all the available unimpaired speakers are considered. The unimpaired speaker who had highest number of closest phones based on percentage overlap (discussed in Step 5) between the formant distributions with the dysarthric speaker is chosen as the closest unimpaired speaker. For Nemours database normal speakers from dialect 2 of TIMIT speech corpus (as dysarthric speakers of Nemours and dialect 2 of TIMIT belong to the North American region they both are expected to share the same dialect) and for Tamil dysarthric speech corpus unimpaired speakers from the same corpus are chosen for analysis.

Step 2 (Formant Frequency Extraction): Next, formant frequencies (F1, F2, and F3) are estimated frame-wise, from all examples of the erroneous phone for the dysarthric speaker and all examples of all phones for the unimpaired speaker. A linear prediction-based formant extraction method is used, with a linear predictor of order 20. To handle the ambiguity that arises in estimating formant frequencies, due to spurious peaks in the

LP spectrum, frequencies that are consistent across all frames of all examples of each phone are considered.

Step 3 (Computing Formant Gaussian Distributions): With the extracted formants, formant Gaussian distributions $N_{Fi_s}(\mu_{Fi_s}, \sigma^2_{Fi_s})$ are derived for the dysarthric and the unimpaired speaker, where Fi denotes the i^{th} formant frequency and s denotes the identity of the speaker.

Step 4 (Computing the Product of Formant Gaussians (PoG)): Using the formant Gaussian distributions (from step 3), the PoG is computed with the dysarthric speaker's phone and phones of the corresponding unimpaired speaker. The product of formant Gaussians [21], $N_{du}(\mu_{du}, \sigma^2_{du})$, where d and u represent the dysarthric and unimpaired speaker respectively, is given by,

$$N_{du}(\mu_{du}, \sigma^2_{du}) = N_{Fi_u}(\mu_{Fi_u}, \sigma^2_{Fi_u}).N_{Fi_d}(\mu_{Fi_d}, \sigma^2_{Fi_d}), \quad (1)$$

where $i = 1, 2, 3$

The perceptual analysis revealed that an erroneous vowel is substituted by another vowel and an erroneous consonant is substituted by another consonant or a vowel. Therefore, if the phone considered is a vowel, PoG is computed with all vowels of the unimpaired speaker. On the other hand, if it is a consonant, like in the case considered, the PoG is computed with vowels and consonants of the unimpaired speaker.

Step 5 (Calculating the Amount of Overlap): With the mean, μ_{du}, normalized amount of overlap, $O^N_{Fi_{du}}$, is calculated to observe the acoustic similarity between the erroneous phone of the dysarthric speaker and the phones of the unimpaired speaker. The normalized amount of overlap, between the two formant Gaussians, $O^N_{Fi_{du}}$, is,

$$O^N_{Fi_{du}} = O_{Fi_{du}} \frac{\sigma_{Fi_d}}{\sigma_{Fi_u}} \quad (2)$$

where the amount of overlap, $O_{Fi_{du}}$, between the two formant Gaussian distributions is given by,

$$O_{Fi_{du}} = \frac{\sigma_{Fi_u}}{\sigma_{Fi_d}} e^{-\left[\frac{(\mu_{du}-\mu_{Fi_u})^2}{2\sigma^2_{Fi_u}} + \frac{(\mu_{du}-\mu_{Fi_d})^2}{2\sigma^2_{Fi_d}}\right]} \quad (3)$$

where $i = 1, 2, 3$

Step 6 (Formant Overlap Threshold Condition): Finally, a threshold for the amount of overlap is chosen empirically to identify the acoustically similar phones. The threshold is given by,

$$O^N_{F1_{du}} \wedge O^N_{F2_{du}} \wedge O^N_{F3_{du}} >= 90\% \quad (4)$$

If the amount of overlap exceeds the threshold, the two phones are considered to be acoustically similar. For example, in Fig. 2 (a), for all three formants, the Gaussians satisfy the overlap threshold condition and hence the two phones of the unimpaired speaker are considered to be acoustically similar to the phone of the dysarthric speaker. On the other hand, in Fig. 2 (b), not all formants satisfy the overlap threshold, and so these phones are acoustically dissimilar and less likely to be confused with each other.

Steps 1 to 6 are repeated for all phones that exhibit substitution errors, for all the dysarthric speakers in both corpora.

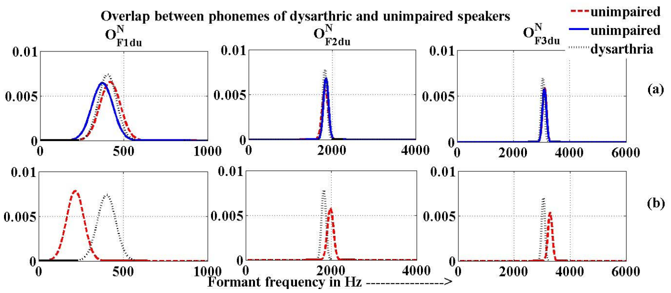

Fig. 2. Amount of overlap ($O^N_{Fi_{du}}$) between formant Gaussians of /ʧ/ of a severe dysarthric speaker and (a) /k/ and /g/, (b) /m/ of an unimpaired speaker.

For the phone considered, /θ/, the phones that satisfy the formant threshold condition are shown in Table I. It is observed that some of these phones are identified as errors in the perceptual analysis as well. The remaining phones that also satisfy the threshold may indicate the recognition errors. The significance of these phones will be discussed further in the upcoming third-level of analysis, which is an acoustic modeling analysis, to further validate and identify the consistency in errors, and also to identify the modeling or recognition errors.

C. ASR-Based Acoustic Modeling Analysis

For this analysis, isolated style (phones are trained and tested in isolation) context-independent and context-dependent speech recognition systems are initially trained. For analysis on the Tamil dysarthric corpus, with both recognition systems, 65% of the analysis data is used for training and the remaining for evaluation. For the Nemours dysarthric speech corpus, due to data insufficiency, speaker-independent models are trained using the data from unimpaired male speakers of the DR2 region in the TIMIT speech corpus [13] and the 65% of the analysis data from the Nemours database is used to adapt models to each dysarthric speaker.

1) Isolated Style Context-Independent Analysis: Initially, Mel frequency cepstral coefficients (MFCC) are extracted from the training data for both Tamil and TIMIT speech corpora. Context-independent HMMs (speaker-dependent for Tamil and speaker-independent for TIMIT) are then trained with 3 states and a varying number of mixture components per state, based on the number of examples for each phone. The trained TIMIT context-independent HMMs are then adapted to each dysarthric speaker in the Nemours dysarthric speech corpus, using maximum a posteriori (MAP) adaptation algorithm. Then, the phones exhibiting substitution error are tested against their acoustically similar phone models, derived from the PoG-based analysis, to validate the previously obtained articulatory errors. The remaining phones are tested against all other phone models to identify the recognition errors.

2) Isolated Style Context-Dependent Analysis: A context-dependent-based isolated style speech recognition system is also trained to analyze the errors. A decision tree-based clustering is used to handle the unseen context-dependent HMMs. For Nemours, 255 context-dependent HMMs common to both TIMIT and Nemours are trained using the TIMIT speech corpus and adapted to each dysarthric speaker of the

TABLE I

AN ILLUSTRATION OF THE SUBSTITUTED PHONE LIST FOR TAMIL AND NEMOURS DYSARTHRIC SPEAKERS IN MILD, MODERATE AND, SEVERE CLASS USING PERCEPTUAL, PRODUCT OF GAUSSIANS BASED FORMANT FREQUENCY AND, AUTOMATIC SPEECH RECOGNITION ANALYSIS

Tamil dysarthric speech corpus											
MRA- mild dysarthric speaker				MGN - moderate dysarthric speaker				MRI - severe dysarthric speaker			
Phone	Substituted phones			Phone	Substituted phones			Phone	Substituted phones		
	Perceptual	PoG	ASR		Perceptual	PoG	ASR		Perceptual	PoG	ASR
/ð/	**/l/, /ð/, θ/**	**/ð/, /l/** /d/, **θ/,** /ɭ/ /t/,	**/ð/, /θ/,** /d/, **/l/**	/a/	**/a/, /aː/**	**/a/, /aː/,** /oː/, /ɯ/	**/a/, /aː/**	/i/	**/i/, /iː/, /e/**	/aː/, /eː/, **/iː/,** /a/, /e/ **/i/,**	/a/, **/i/, /iː/, /e/**
/ʃ/	**/s/, /ʃ/**	**/s/, /ʃ/**	**/s/, /ʃ/**	/b/	**/p/, /u/, /g/, /ð/,** /k/	**/ð/, /u/,** /uː/, **/g/,** /y/, /r/, **/p/**	**/p/, /u/, /g/, /ð/ ,** /d/	/θ/	**/k/, /θ/**	**/θ/,** /d/, /e/, **/k/,** /g/, /h/, /l/, /t/, /uː/, /ɭ/, /u/	**/k/,** /t/, /g/, /h/, **/θ/,** /e/, /l/
Nemours dysarthric speech corpus											
MH - mild dysarthric speaker				JF - moderate dysarthric speaker				SC - severe dysarthric speaker			
/ʧ/	**/ʧ/,** /ʃ/	**/ʧ/**	**/ʧ/**	/t/	**/t/, /n/**	/d/, **/t/,** /ʧ/, /p/ /f/, /g/, /h/, **/n/,** /iː/, /ʤ/, /k/, /ʒ/	/d/, **/t/,** /h/, /iː/, /k/, /p/, /s/, /j/, **/n/**	/iː/	**/i/, /iː/,** /ɛ/	**/i/, /iː/**, /ɛ/	**/i/, /iː/**
/f/	**/eɪ/,** /θ/, **/f/**, /ð/	**/eɪ/,** /ɪ/, **/f/,** /v/, /p/, /b/, /d/	**/eɪ/,** /ɪ/, **/f/,** /b/, /d/	/w/	**/w/, /n/**	**/w/,** /b/, /d/, **/n/,** /ng/	**/n/,** /d/ /b/, **/w/**	/s/	**/s/, /z/,** /f/, **/ʤ/,** /r/	/ʧ/, /g/, **/s/ /ʤ/,** /h/, /k/, /p/, /ʃ/, **/z/,** /ʒ/	/ʧ/, **/ʤ/,** /h/ /ʃ/, **/z/ /s/,** /p/

Nemours database. For the Tamil dysarthric speech corpus, 1741 context-dependent HMMs are trained. This system is also tested similar to the context-independent system and it is observed that there is a reduction in the number of modeling errors with the context-dependent system, when compared to the context-independent system, due to the addition of contextual information.

Table I shows the list of substituted phones obtained through perceptual, PoG-based acoustic similarity, and isolated style ASR analyses. The list of phones that are found to be common across all the three analyses are considered to be substituted phones due to speaker-specific articulation or pronunciation errors. The phones identified to be erroneous in both, the PoG and the isolated style ASR analyses, are considered to exhibit acoustic modeling error.

The benefits of distinguishing between articulatory and modeling errors are two fold. (i) The knowledge of speaker-specific articulatory errors would aid in providing appropriate speech therapy to dysarthric speakers based on their errors. (ii) As the modeling accuracy is improved, the number of acoustic modeling errors could be reduced to a greater extent, while the speaker-specific articulatory errors would still prevail due to the neuro-physiological conditions of the speakers. If the errors are not distinguished, then, when the modeling accuracy is improved, the entire error analysis and the weight computation process (described below) would have to be repeated. However, distinguishing the errors would make allowances for the analysis and modification of the modeling errors alone.

In order to correct articulatory and modeling errors, a speaker-specific phone confusion transducer for both context-independent and context-dependent systems would have to be developed, for which weights are to be computed for each phone. These weights are computed based on the isolated style recognition analyses, as described below.

Weight Calculation: Weights are calculated for both articulatory errors and recognition errors, based on the performance of the recognition system, for all 30 dysarthric speakers, as follows:

$$w = 1 - p_{ij}^{k} \tag{5}$$

where p_{ij}^{k} refers to probability that phone j is substituted by phone i, for dysarthric speaker, k, and is given by

$$p_{ij}^{k} = \frac{number\ of\ occurrences\ of\ phone\ i}{total\ number\ of\ examples\ of\ phone\ j} \tag{6}$$

The WFST is designed to choose the lowest path. Weights for each substituted phone, of each erroneous phone, are calculated based on both context-independent and context-dependent analysis. These weights are then used to create context-independent and context-dependent phone confusion transducers to correct errors that appear in the text derived from the speech recognition system of an AASC aid. The details of the AASC aid are elaborated in the following section.

IV. AUGMENTATIVE AND ALTERNATIVE SPEECH COMMUNICATION AID

An augmentative and alternative speech communication (AASC) aid consists of a speaker-dependent dysarthric speech recognition (DSR) system, an error-correction system, and a text-to-speech synthesis system, as shown in Fig. 3. Dysarthric speech is initially recognized by a speaker-dependent continuous DSR system. The erroneous text from the DSR system is then corrected using a speaker-specific weighted finite state transducer (WFST), whose weights are computed based on the speaker-specific error analysis, discussed in Section III. Finally, the error-corrected text is synthesized using a text-to-speech synthesis system. The components of the AASC system are described as follows:

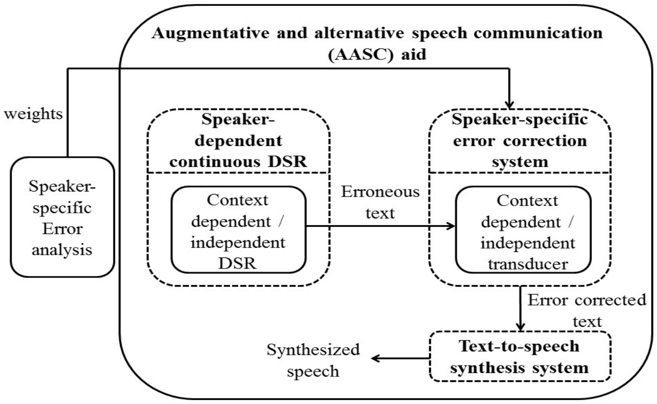
Augmentative and alternative speech communication (AASC) aid
weights
Speaker-specific Error analysis
Speaker-dependent continuous DSR
Context dependent / independent DSR
Erroneous text
Speaker-specific error correction system
Context dependent / independent transducer
Error corrected text
Text-to-speech synthesis system
Synthesized speech

Fig. 3. Block diagram for augmentative and alternative speech communication aid.

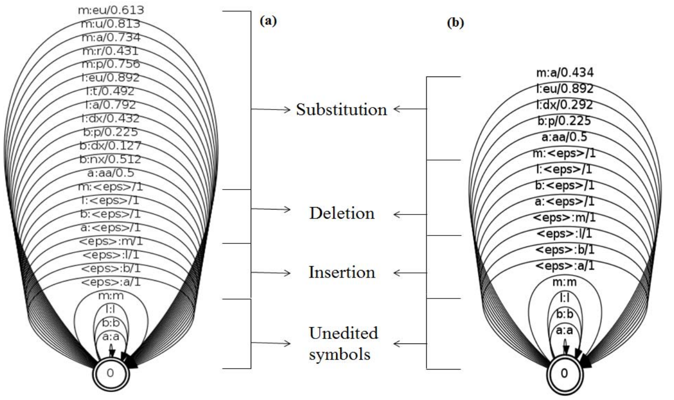

Fig. 4. Phone confusion transducer for the word “balam” of a Tamil moderate dysarthric speaker (a) context-independent level and (b) context-dependent level.

A. Speaker-Dependent Continuous Dysarthric Speech Recognition (DSR) Systems

The first component involved in an AASC aid is a DSR system. Two continuous dysarthric speaker-dependent DSR systems, namely (i) a context-independent (CI) DSR system, with phone-level bigram statistics, and (ii) a context-dependent (CD) DSR system, are trained for the 20 dysarthric speakers in the Tamil dysarthric speech corpus. For the Nemours dysarthric speech corpus, owing to the limited amount of data available, speaker-adaptive DSR systems, namely (i) a CI system with phone-level bigram statistics, and (ii) a CD system are developed.

For training, 200 sentences from the Tamil dysarthric speech corpus and 54 sentences from the Nemours dysarthric speech corpus are used for each dysarthric speaker. For testing, remaining 165 Tamil utterances and 20 Nemours utterances are used. Performances of these speech recognition systems, for each dysarthric speaker, is evaluated through phone error rate (PER), in percentage, given in equation 7:

$$PER = \frac{S + D + I}{N} * 100\% \tag{7}$$

where, N is the total number of phones, and D, S, and I are the number of deleted, substituted, and inserted phones, respectively.

A description of the continuous speech recognition systems, trained for the two corpora, is as follows:

1) Tamil Dysarthric Speech Corpus:

- **Context-independent (CI) system** For the speaker-dependent CI DSR system, CI HMMs are trained, with the dysarthric speakers' data, with 3 states and a varying number of mixture components per state, based on the number of examples for each phone. Phone-level bigram statistics are derived and a decoding network is created using these bigram statistics [23]. Each speaker-dependent system is tested and the recognized text is evaluated using equation 7. The performance of this system is shown in Table II (column 2 and 9) for all the 20 dysarthric speakers.
- **Context-dependent (CD) system** Context-dependent HMMs are trained, with 3 states and 5 mixture components per state, to include the contextual information. Number of models in the system is as mentioned in Section III-C. Performance of this system is shown in column 4 and 11 of Table II and it is observed that the PER obtained is less than that obtained with the CI system by up to 27.72%.

2) Nemours Dysarthric Speech Corpus:

- **Context-independent (CI) speaker adaptive system** For the CI speaker-adaptive system, unimpaired male speakers from the TIMIT speech corpus (DR2) are chosen for training speaker-independent context-independent models. These context-independent models are then adapted to each dysarthric speaker using MAP adaptation. Phone-level bigram statistics are used here too. For adaptation, 54 utterances from each of the dysarthric speakers in the Nemours database are used. Performance of the system is tabulated in Table II (column 9).
- **Context-dependent (CD) speaker adaptive system** Speaker-independent CD models are trained as discussed in Section IV-A.1, for the TIMIT speech corpus (DR2) and these are adapted to the Nemours dysarthric speakers, individually. This system reduces the PER by up to 16.26%, when compared with the speaker-adaptive CI system (refer to Table II (column 11)).

From the above discussions, it is observed that the errors due to acoustic modeling are reduced to a greater extent in the CD DSR systems, as reflected by the PERs in Table II. To improve the performance of these systems further, observations from the error analysis, discussed in Section III, have to be incorporated, as discussed below.

B. Speaker-Specific Error Correction System

The errors of the DSR system can be reduced by means of a speaker-specific error correction system, which makes use of a phone confusion transducer, created using weighted finite state transducers (WFSTs). In the current work, input symbols in the WFST refer to the erroneous phones, output symbols refer to their acoustically similar phones, and weights refer to the weights calculated in Section III-C. Steps involved in developing the transducer and correcting the errors are as follows:

TABLE II

DSR PERFORMANCE IN %PER AND %WER FOR BASELINE DSR AND WFST POST-PROCESSED DSR SYSTEMS (T - TAMIL DYSARTHRIC SPEAKERS; N - NEMOURS DYSARTHRIC SPEAKERS; M - MILD; MOD - MODERATE; S - SEVERE DYSARTHRIC SPEAKERS)

Speaker id	PER				WER		Speaker id	PER				WER	
	CI		CD		Bigram WFST			CI		CD		Bigram WFST	
	Baseline	WFST	Baseline	WFST	CI	CD		Baseline	WFST	Baseline	WFST	CI	CD
FVP (T-M)	33.6	23.72	5.88	4.69	**18.21**	**16.86**	MSU (T-Mod)	54.51	40.78	51.42	43.59	**44.97**	**40.21**
MRA (T-M)	40.43	28.69	32.06	20.37	**31.46**	**29.26**	FBL (T-Mod)	56.43	44.51	55.27	40.51	**45.21**	**43.11**
MPR (T-M)	38.08	21.68	30.23	20.71	**23.42**	**20.78**	MMA (T-S)	62.63	45.47	59.69	42.12	**47.92**	**46.29**
MPA (T-M)	44.43	40.29	39.61	34.7	**16.33**	**9.21**	MRI (T-S)	65.44	54.68	63.09	51.75	**46.41**	**45.53**
MPK (T-M)	43.95	33.7	37.25	35.51	**28.92**	**23.68**	MER (T-S)	68.8	60.69	63.25	50.69	**49.89**	**49.34**
FSI (T-M)	43.95	36.28	12.16	10.21	**18.21**	**10.39**	FB (N-M)	34.15	21.65	30.3	20.17	**28.74**	**26.67**
MAK (T-M)	37.6	35.41	35.09	32.34	**28.86**	**19.08**	BB (N-M)	41.74	24.92	36.85	16.82	**23.41**	**20.83**
MKA (T-Mod)	46.19	25.77	45.19	24.19	**37.91**	**37.85**	MH (N-M)	46.95	20.73	44.27	13.41	**28.03**	**25**
MVI (T-Mod)	53.44	35.08	51.72	27.59	**43.98**	**43.15**	LL (N-M)	42.25	28.7	39.38	27.33	**38.14**	**35.83**
FGA (T-Mod)	55.04	42.88	51.35	38.69	**45.51**	**42.2**	RL (N-Mod)	65.15	45.76	61.18	28.79	**38.83**	**35.83**
MGN (T-Mod)	55.79	37.68	52.6	34.66	**44.21**	**43.21**	JF (N-Mod)	63.55	30.89	47.29	26.65	**39.21**	**39.17**
FAM (T-Mod)	46.24	41.21	40.07	38.92	**27.31**	**22.37**	BV (N-Mod)	74.03	51.34	65.52	46.57	**44.72**	**44.17**
FSP(T-Mod)	44.8	38.28	41.25	37.38	**20.53**	**13.16**	SC (N-S)	59.33	52.39	56.45	33.62	**32.83**	**30.83**
FDH (T-Mod)	50.24	43.54	46.31	42.43	**38.96**	**30.92**	BK (N-S)	81.21	51.21	70.91	48.03	**42.23**	**41.67**
MMU (T-Mod)	52.75	44.27	51.58	40.25	**29.31**	**26.97**	RK (N-S)	72.59	53.17	65.56	44.29	**33.71**	**32.5**
Mean								**52.5**	**38.5**	**46.1**	**32.5**	**34.6**	**31.5**

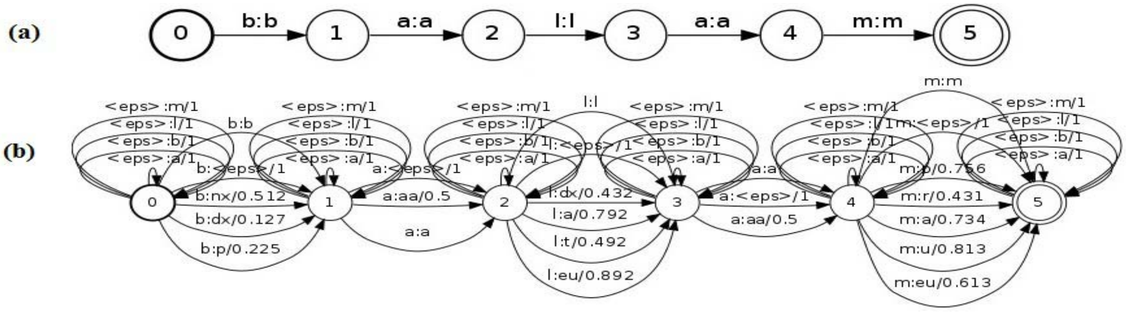

Fig. 5. WFST (a) Reference transducer, (b) Composed transducer.

Step 1 (Phone Confusion Transducer): A phone confusion transducer must correct the insertion, deletion, and substitution errors that are expected to appear in a recognized text, leaving the phones (symbols) that are recognized correctly unchanged. To correct substitution errors, if input symbol 'x' is known to be substituted by 'y', the WFST is represented as 'y:x/w', where 'w' is the corresponding weight [24]. For the insertion of a symbol 'x', the WFST is represented as '<eps>:x/w' and the deletion of a symbol 'x' is represented as 'x:<eps>/w'. To handle the correctly recognized symbols, the WFST is represented as 'x:x/w'.

The WFST is designed to derive the output text based on the lowest weight (minimum edit distance algorithm) associated with it. A phone might be recognized correctly or erroneously. Therefore, for including the chances of its correct recognition, a minimum weight of zero is assigned to all the phones. For the least dominant errors, namely for insertions and deletions, (as discussed in Section III-A) highest weight of 1 is assigned. For substitution errors, the weights computed in Section III-C are used.

For each of the dysarthric speakers (20 Tamil and 10 Nemours), a separate speaker-specific phone confusion transducer is developed. Phone confusion transducers for the context-independent and context-dependent-based DSR systems, discussed earlier, are developed for comparison.

As an example, the phone confusion transducers for the word "balam" (meaning strength) (containing the phones, /a/, /b/, /l/, /m/) are shown in Fig. 4. It can be observed that the number of confusions in the context-dependent transducer in Fig. 4 (b) is less than those in the context-independent transducer in Fig. 4 (a).

Step 2 (Reference Transducer): Word-reference transducers are created for all the unique words in the speech corpus, as shown in Fig. 5 (a). If vocabulary size is to be increased later, only the word-reference transducers have to be added to the available ones, while the speaker-specific confusion transducers, derived in Step 1, need not be modified.

Step 3 (Composition Transducer): The phone confusion transducer in Fig. 4 (a) and the reference transducer in Fig. 5 (a) are to be composed, as shown in Fig. 5 (b).

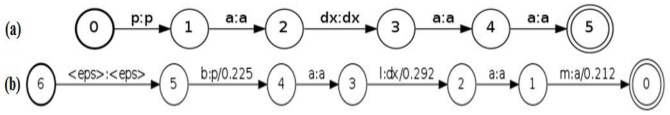

Fig. 6. (a) Test FST for the word "balam" from the DSR output and (b) Output FST after minimum edit distance algorithm.

Step 4 (Union Transducer): The composed transducers are then converted to a single union transducer. For Tamil, it is a union of 780 composed transducers and for Nemours, it is 112 composed transducers.

Step 5 (Testing): The recognized text from the various DSR systems discussed in Section IV-A are given as test FSTs, as shown in Fig. 6 (a). The errors (insertions, deletions, and substitutions) in the test FSTs are corrected by computing minimum edit distance between the test FSTs and the union WFST, to obtain the actual word in the output FST (refer Fig. 6 (b)), using equation 8.

$$W_c = argmin[E_{dist}(P_1 \cup P_2 \cup P_3 \cup \cdots \cup P_n)] \qquad (8)$$

Here W_c refers to the corrected output word and P_1, $P_2, \ldots, P_n$ refer to the n (780 for the Tamil corpus and 112 in the Nemours corpus) entities (composition transducers) in the union WFST. The output from the FSTs (refer Fig. 6 (b)) are evaluated in terms of PER, using equation 7.

1) Results and Discussions:

Table II (columns 3, 5, 10, 12) shows recognition performance after correcting the errors using WFST for the recognized text from the various speaker-dependent speech recognition systems trained for both the speech corpora. Performance is measured based on the phone error rate (PER), given in equation 7.

The current work has reduced the PER by up to 11.69% for mild, 24.13% for moderate, and 12.56% for severe dysarthric speakers, when compared to that of the baseline DSR systems, for the Tamil dysarthric speech corpus. The proposed system gives a reduced PER for moderate dysarthric speakers when compared to mild and severe. Presumably, it is because severe speakers are too unintelligible to improve much, and mild dysarthric speakers are already quite intelligible to improve further. In Nemours, for mild speakers, the reduction in PER, compared to the baseline DSR system, is up to 30.86%, for moderate, 32.39%, and for severe dysarthric speakers, 22.88%. It is observed that the error rates obtained for the Tamil corpus are less than those obtained for the Nemours database. This could be owing to the use of speaker-dependent systems for the former corpus and speaker-adaptive ones for the latter. The greater improvement in performance, when compared with the baseline system, with the Nemours database, than with the Tamil database, can be attributed to the restricted syntax of the utterances in the former corpus.

The performance in Table II is in terms of the number of phones correctly or incorrectly recognized. However, from the error-corrected text, three possibilities can be observed, namely (i) incorrect, (ii) partially correct, and (iii) completely correct utterances. To improve the performance further, from the partially correct utterances, a word bigram FST is constructed for Nemours and the Tamil corpus individually. This bigram FST is composed with the union FST discussed in Section IV-B and it is tested with the test FST. Table II shows the results obtained after composing with word bigram FST for both context-independent and context-dependent systems, and the performance is measured in terms of word error rate (WER).

The performance of the current work is compared with [14] (described in Section I). Here, the Nemours dysarthric speech corpus is used and the WER of a context-independent speech recognition system, with WFST-based post-processing, for low intelligibility speakers (RK, RL, BV, BK, SC) is mentioned to be up to 48% and for high intelligibility speakers (BB, FB, JF, LL, MH), up to 26%. In the current work, (refer to Table II column 13) the context-independent system results in a WER that is as low as 32.83% and 23.41%, for low and high intelligibility speakers, respectively. The context-dependent system further reduces the WER to 30.83% and 20.83% (refer Table II column 14), which is 17.17% and 5.17% lower than that obtained in [14], for low and high intelligibility speakers, respectively.

An AASC aid is complete only when the recognized text is synthesized as speech that is intelligible compared to the original dysarthric speech. The text-to-speech synthesis system used to accomplish this is described below.

C. Text-to-Speech Synthesis System

An HMM-based text-to-speech synthesis system (HTS) is used to synthesize the error corrected text from the WFST. Since the Tamil dysarthric speech corpus has both male and female dysarthric speakers and Nemours has only male dysarthric speakers, 3 TTS synthesizers are built, one each for Tamil male, Tamil female, and English male speakers.

For the training phase in Tamil TTS systems, three hours of Tamil speech data [19] each from an unimpaired male and female speaker is used. For Nemours, one hour of speech data each, from 3 male speakers of the CMU Arctic database [25] is used. HMMs (speaker-dependent for Tamil and speaker-independent for English) with 5 states and a single mixture component per state are trained with 108-dimensional feature vectors, consisting of Mel generalized cepstral coefficients and log fundamental frequency. The error-corrected texts, corresponding to all the test utterances, derived from the WFST, are synthesized.

The synthesized speech signals, for both dysarthric speech corpora, are tested for intelligibility, through a listening test. For the test, 25 utterances per dysarthric speaker are evaluated by 37 expert listeners, whose ages range between 23 and 35 years, in a laboratory environment. Each expert assesses 6 dysarthric speakers' natural and synthesized speech data, such that all the 30 dysarthric speakers are assessed by at most 7 different listeners. The listeners are asked to write the text they perceive for both the natural dysarthric and the synthesized utterances. A WER is then computed manually, for natural and synthesized dysarthric speech. Table III shows results of the intelligibility analysis and it is interesting to note that this technique of correcting error through phonological analysis has improved the intelligibility score for severe dysarthric speakers by reducing the WER to up to 41.52%, from 100%. The agreement between the listeners is found to

TABLE III
WORD ERROR RATE (WER) FROM SUBJECTIVE INTELLIGIBILITY TEST (T - TAMIL DYSARTHRIC SPEAKERS; N - NEMOURS DYSARTHRIC SPEAKERS; M - MILD; MOD - MODERATE; S - SEVERE DYSARTHRIC SPEAKERS

Speaker id	Natural	Syn	Speaker id	Natural	Syn
FVP (T-M)	9.88	8.55	MSU (T-Mod)	35.31	22.13
MRA (T-M)	9.9	9.59	FBL (T-Mod)	39.14	26.59
MPR (T-M)	11.32	10.77	MMA (T-S)	100	41.52
MPA (T-M)	12.23	9.12	MRI (T-S)	100	45.88
MPK (T-M)	17.76	15.71	MER (T-S)	100	45.77
FSI (T-M)	11.83	11.42	FB (N-M)	17.32	8.23
MAK (T-M)	14.9	12.29	BB (N-M)	9.91	9.75
MKA (T-Mod)	73.4	28.69	MH (N-M)	11.76	9.45
MVI (T-Mod)	71.32	26.59	LL (N-M)	30.31	13.69
FGA (T-Mod)	63.12	21.87	RL (N-Mod)	34.21	13.79
MGN (T-Mod)	83.13	27.87	JF (N-Mod)	67.12	28.58
FAM (T-Mod)	35.31	15.65	BV (N-Mod)	41.23	28.69
FSP(T-Mod)	41.93	23.46	SC (N-S)	100	31.2
FDH (T-Mod)	36.89	31.32	BK (N-S)	100	45.88
MMU (T-Mod)	53.12	40.23	RK (N-S)	100	42.12
Mean				**58.46**	**27.35**

have a kappa value of 0.435, meaning, there is a moderate agreement between the listeners.

V. CONCLUSIONS

The current work focuses on the development of an AASC aid for dysarthric speakers. Major issue in the development of any AASC aid is in correcting the errors in the text derived from the DSR system. These errors may occur owing to inability of the dysarthric speaker to utter certain phones correctly (articulatory or pronunciation error) or due to inaccurate models in the recognition system (modeling error). While existing approaches mostly focus on eliminating the articulatory errors, the current work focuses on handling modeling errors as well. Further, even though certain techniques do attempt to handle both errors, they do not differentiate between them, unlike in the current work. Also, the analyses in the current work are not restricted to a fixed syntax of sentences. In order to identify and distinguish between the two errors, a three-level cascaded analysis is performed. Using the results of this analysis, phone confusion transducers are trained for each dysarthric speaker, to correct the errors in the text derived from the recognition system. The corrected text is finally synthesized by HMM-based TTS systems to yield intelligible speech.

Though the proposed AASC technique is observed to improve the intelligibility of dysarthric speech considerably in all the three category of dysarthric speakers it also faces few challenges. A substantial amount of time is required for dysarthric speech data collection and phone-level annotation. To overcome this issue we would like to collect speech data with multiple examples for words using array microphones and apply multi-resolution features that are expected to improve the performance further. Though AASC makes use of PoG and ASR-based acoustic analysis to validate the results obtained through perceptual analysis a complete automated technique is required to differentiate articulatory and modeling errors. Based on the feedback from the AASC users the authors would like to build AASC aids that are specific to their occupation and independent living.

ACKNOWLEDGEMENTS

The authors would like to thank the National Institute of Empowerment of Persons with Multiple Disabilities (NIEPMD), for their constant support in collecting dysarthric speech data.

REFERENCES

[1] R. Mukherjee, S. Dey, S. Das, and A. Basu, "An iconic and keyboard based communication tool for people with multiple disabilities," in *Proc. IEEE Students Technol. Symp. (TechSym)*, Apr. 2010, pp. 282–288.

[2] J. Murphy, "I prefer contact this close': Perceptions of AAC by people with motor neurone disease and their communication partners," *Augmentative Alternative Commun.*, vol. 20, no. 4, pp. 259–271, 2004.

[3] M. S. Hawley *et al.*, "A voice-input voice-output communication aid for people with severe speech impairment," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 21, no. 1, pp. 23–31, Jan. 2013.

[4] P. D. Polur and G. E. Miller, "Investigation of an HMM/ANN hybrid structure in pattern recognition application using cepstral analysis of dysarthric (distorted) speech signals," *Med. Eng. Phys.*, vol. 28, no. 8, pp. 741–748, Oct. 2006.

[5] F. Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 947–960, May 2011.

[6] M. S. Hawley *et al.*, "A speech-controlled environmental control system for people with severe dysarthria," *Med. Eng. Phys.*, vol. 29, no. 5, pp. 586–593, Jun. 2007.

[7] M. Hasegawa-Johnson, J. Gunderson, A. Perlman, and T. Huang, "Hmm-based and Svm-based recognition of the speech of talkers with spastic dysarthria," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2006, pp. 1060–1063.

[8] M. Kim, Y. Kim, J. Yoo, J. Wang, and H. Kim, "Regularized speaker adaptation of KL-HMM for dysarthric speech recognition," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 9, pp. 1581–1591, Sep. 2017.

[9] S. R. Shahamiri and S. S. B. Salim, "A multi-views multi-learners approach towards dysarthric speech recognition using multi-nets artificial neural networks," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 5, pp. 1053–1063, Sep. 2014.

[10] H. V. Sharma and M. Hasegawa-Johnson, "Acoustic model adaptation using in-domain background models for dysarthric speech recognition," *Comput. Speech Lang.*, vol. 27, no. 6, pp. 1147–1162, Sep. 2013.

[11] K. T. Mengistu, F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2011, pp. 4924–4927.

[12] M. J. Kim, Y. Kim, and H. Kim, "Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, pp. 694–704, Apr. 2015.

[13] M. Dhanalakshmi, T. A. Mariya Celin, T. Nagarajan, and P. Vijayalakshmi, "Speech-input speech-output communication for dysarthric speakers using HMM-based speech recognition and adaptive synthesis system," *Circuits, Syst., Signal Process.*, vol. 37, no. 2, pp. 647–703, Feb. 2018.

[14] S. O. Caballero Morales and S. J. Cox, "Modelling errors in automatic speech recognition for dysarthric speakers," *EURASIP J. Adv. Signal Process.*, vol. 1, pp. 308–340, Dec. 2009.

[15] X. Menendez Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The Nemours database of dysarthric speech," in *Proc. 4th Int. Conf. Spoken Lang. Process.*, vol. 3, Oct. 1996, pp. 1962–1965.

[16] R. D. Kent, G. Weismer, J. F. Kent, J. K. Vorperian, and J. R. Duffy, "Acoustic studies of dysarthric speech: Methods, progress, and potential," *J. Commun. Disorders*, vol. 32, no. 3, pp. 141–186, May/Jun. 1999.

[17] T. A. Mariya Celin, T. Nagarajan, and P. Vijayalakshmi, "Dysarthric speech corpus in tamil for rehabilitation research," in *Proc. IEEE Region 10th Conf. TENCON*, Nov. 2016, pp. 2610–2613.

[18] W. M. Fisher and G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proc. DARPA Workshop Speech Recognit.*, 1986, pp. 93–99.

[19] G. Anushiya Rachel, V. Sherlin Solomi, K. Naveenkumar, P. Vijayalakshmi, and T. Nagarajan, “A small-footprint context-independent HMM-based synthesizer for tamil,” *Int. J. Speech Technol.*, vol. 18, no. 3, pp. 405–418, Sep. 2015.

[20] S. Manochiopinig, N. Thubthong, and, P. Kayasith, “Dysarthric speech characteristics of thai stroke patients,” *Disab. Rehabil. Assistive Technol.*, vol. 3, no. 6, pp. 332–338, Nov. 2008.

[21] T. Nagarajan and D. O’Shaughnessy, “Bias estimation and correction in a classifier using product of likelihood-Gaussians,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 3, Apr. 2007, pp. III-1061–III-1064.

[22] S. S. VijayaRajSolomon, V. Parthasarathy, and N. Thangavelu, “Exploiting acoustic similarities between Tamil and Indian English in the development of an HMM-based bilingual synthesiser,” *IET Signal Process.* vol. 11, no. 3, pp. 332–340, May 2017.

[23] S. Young *et al.*, *The HTK Book (for HTK Version 3.4)*. Cambridge, U.K.: Cambridge Univ. Dept. Eng., 2002.

[24] M. Mohri, F. Pereira, and M. Riley, “Weighted finite-state transducers in speech recognition,” *Comput. Speech Lang.*, vol. 16, no. 1, pp. 69–88, Jan. 2002.

[25] J. Kominek and A. W. Black, “The CMU Arctic speech databases,” in *Proc. 5th ISCA Speech Synth. Workshop*, Pittsburgh, PA, USA, Jun. 2004, pp. 223–224.