iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** VII **Month of publication:** July 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Review on Speech Assistive Technologies for Dysarthric Patients

Lovee Jain[1], Prema N[2], Vedavathi N[3]
Department of Computer Science and Engineering, NIE Institute of Technology, Mysore, Karnataka, India

Abstract- Over the past decade, several speech based electronic assistive technologies have been developed that target users with dysarthric speech. These devices include vocal command and control systems, but also voice input voice output communication aids. This paper aims to survey on various speech assistive technologies which can be used by dysarthric patients and helps them in communication. It also describes the various types of dysarthria on which these devices are based and the challenges which are faced while creating these assistive devices.
Keywords- ASR, TTS, concatenation technique, grafting and TORGO Morph and ALADIN

I. INTRODUCTION

Speech holds an important role in the evolution of human civilization not only being the spoken form of a language but also the most efficient way of communication. Spoken language communication is central to daily life, but as many as 1.3% of the population cannot use natural speech to communicate reliably [1,2]. Dysarthria [1,2] is a motor speech disorder affecting millions of people. A dysarthric speaker has much difficulty in communicating. This disorder induces bad or not pronounced phonemes, variable speech amplitude, poor articulation etc. In other words, it is a condition in which problems effectively occur with the muscles that help produce speech, often making it very difficult to pronounce words. The type and severity of dysarthria depend on which area of the nervous system is affected. Several authors have classified the types of dysarthria taking into consideration the symptoms of neurological disorder[1]. All types of dysarthria affect the articulation of consonants, causing the slurring of speech. In very severe cases, vowels may also be distorted. Intelligibility varies greatly depending on the extent of neurological damage. The general types of dysarthria can be considered as follows:

A. Spastic Dysarthria

This is due to exaggerated stretch reflexes, resulting in increased muscle tone and in coordination. Vocal quality is harsh. Sometimes the voice of a patient with spastic dysarthria is described as strained or strangled. Pitch is low, with pitch breaks occurring in some cases. Bursts of loudness are sometimes noted.

B. Hyperkinetic Dysarthria

Hyperkinetic dysarthria is usually thought to be due to lesions of the basal ganglia. Its predominant symptoms are associated with involuntary movement. Vocal quality may be described as harsh, strained, or strangled. Voice stoppages may occur in dysarthria associated with dystonia.

C. Hypokinetic Dysarthria

This is associated mainly with Parkinson's disease. Hoarseness is common in Parkinson's patients. Also, low volume frequently reduces intelligibility.

D. Ataxic Dysarthria

This disorder is due to damage to the cerebellar control circuit. It can affect respiration, phonation, resonance and articulation, but its characteristics are most pronounced in articulation and prosody.

E. Flaccid Dysarthria

This result from damage to the lower motor neurons (cranial nerves) involved in speech. Monopitch and monoloudness may both result from vocal fold paralysis.

F. Mixed Dysarthria

Characteristics will vary depending on whether the upper or lower motor neurons remain most intact

Over the past decades, several speech based electronic assistive technologies have been developed that target users with dysarthria. Based on the severity level of dysarthria, different assistive technologies are used but still none of the techniques can guarantee for 100% accuracy and many challenges have been faced in creating these assistive devices. The main three challenges faced by the system are

Dysarthric speech varies greatly between speakers which is mainly based on the severity and type of dysarthria.

Speaking often requires more effort, thus the amount of training or adaption material that can be collected is restricted. Nemours data is usually collected for the same reason but collecting actual data is again a big difficulty.

The number of phones that can be produced is often severely restricted, making it difficult to distinguish between words.

In the following sections, various speech assistive techniques are discussed along with the assessment techniques which can diagnose the degree of speech disorder.

II. SPEECH ASSISTIVE TECHNIQUES FOR DYSARTHRIA

There are many speech assistive techniques which can help dysarthric patients to communicate and to explain their ideas. The conventional and common electronic assistive technologies for dysarthric speech are based on two steps: i) Using an Automatic speech recognition [2] ii) Synthesizing it along with concatenation and grafting the phonemes[2].

A. Automatic Speech Recognition

For an ASR system, a speech signal refers to the analogue electrical representation of the acoustic wave, which is a result of the constrictions in the vocal tract. Different vocal tract constrictions generate different sounds. The basic sound in a speech signal is called a phoneme. These phonemes are then combined, to form words and sentences. Each phoneme is dependent on its context. Each language has its own set of distinctive phonemes, which typically amounts to between 30 and 50 phonemes. An ASR [5, 7] system mainly consists of four components: pre-processing stage, feature

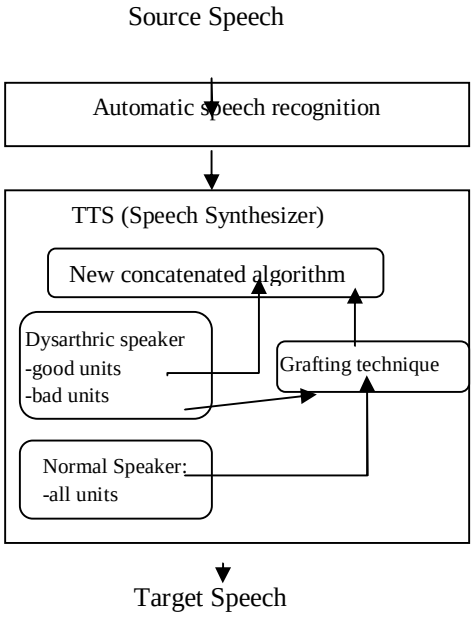


Fig 1. The overall system design to help dysarthric speakers

extraction stage, classification stage and a language model.

1) *Pre-processing:* The pre-processing stage transforms the speech signal before any information is extracted by the feature extraction stage. The functions to be implemented by the pre-processing stage are also dependent on the approach that will be employed at the feature extraction stage. A number of common functions are the noise removal, endpoint detection, pre-emphasis, framing and normalisation.
2) *Feature Extraction*: After pre-processing, the feature extraction stage extracts a number of predefined features from the processed speech signal. These extracted features must be able to discriminate between classes while being robust to any external conditions, such as noise. Therefore, the performance of the ASR system is highly dependent on the feature extraction method chosen, since the classification stage will have to classify efficiently the input speech signal according to these extracted features. Over the past few years various feature extraction methods have been proposed, namely the MFCCs, the discrete wavelet transforms (DWTs) and the linear predictive coding (LPC)
3) *Language model*: The next stage is the language model, which consists of various kinds of knowledge related to a language, such as the syntax and the semantics. A language model is required, when it is necessary to recognise not only the phonemes that make up the input speech signal, but also in moving to either trigram, words or even sentences. Thus, knowledge of a language is necessary in order to produce meaningful representations of the speech signal.

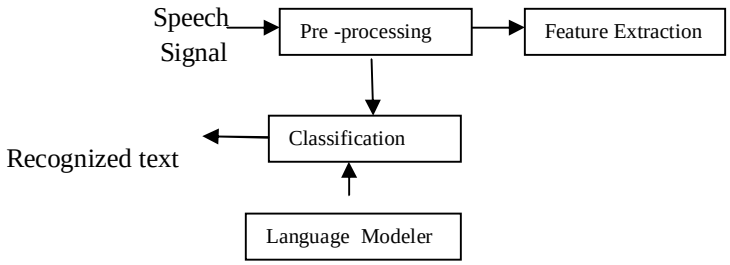


Fig 2. Automatic Speech Recognition System Concept

4) Classification: The final component is the classification stage, where the extracted features and the language model are used to recognise the speech signal. The classification stage can be tackled in two different ways: The first approach is the generative approach, where the joint probability distribution is found over the given observations and the class labels. The resulting joint probability distribution is then used to predict the output for a new input. Two popular methods that are based on the generative approach are the Hidden Markov Models HMMs[6] and the Gaussian mixture models (GMMs). The second approach is called the discriminative approach. A model based on a discriminative approach finds the conditional distribution using a parametric model, where the parameters are determined from a training set consisting of pairs of the input vectors and their corresponding target output vectors. Two popular methods that are based on the discriminative approach are the ANNs and support vector machines (SVMs)

B. Synthesis of Dysarthric Speech

The synthesis of Dysarthric speech can be done by TTS[2,8] conversion, concatenation and by grafting techniques. A text to speech synthesizer is a computer based system that can read text aloud automatically, regardless of whether the text is introduced by a computer input stream or a scanned input submitted to an OCR engine. Here the result from the ASR is taken as an input for the TTS synthesizer. The TTS system comprises of five fundamental components:

1) *Text Analysis And Detection*: the text analysis is a pre-processing part which analyse the input text and organize into manageable list of words.
2) *Text Normalization And Linearization*: The text normalization is transformation of text to pronounceable form.
3) *Phonetic Analysis*: Phonetic analysis converts the orthographical symbols into phonological ones using a phonetic alphabet. Pronunciation of words based on its spelling has two approaches to do speech synthesis namely i) Dictionary based approach ii) Rule based approach
4) *Prosodic Modelling And Intonation*: The prosodic modelling describes the speaker's emotion. The identification of the vocal features which signal emotional content helps to create very natural synthesized speech. Intonation is simply a variation of

speech while speaking.

5) *Acoustic Processing*: The speech will be spoken according to the voice characteristics of a person. There are three types of acoustic synthesizing available i) Concatenative synthesis ii) Formant synthesis iii) articulatory synthesis

The conventional methods, along with ASR and TTS system, use some concatenation and drafting techniques to boost the performance for impaired speech. The concatenation algorithm proposed by Mohammed Sidi Yakcoub *el a*t[1,2] is explained as follows:

C. Concatenation Algorithm

The units will be concatenated at the edge starting or ending of vowels. The algorithm always concatenate two periods of the same vowel with different shapes in time domain. It concatenates /a/ and /a/,/e/ and/e/ etc. The main steps in an algorithm are

1) take one period from the left unit(LP)
2) take one period from the right unit(RP)
3) Use a warping function to convert LP to RP in frequency domain.
4) Each converted period is followed by an interpolation in the time domain.

D. Grafting Technique

Dysarthria affects the pronunciation of specified phonemes depending on the speaker and/or type of dysarthria. Hence to make the speech more intelligible, changes in the words or correction is required which can be done with grafting technique. The main purpose of grafting technique is to remove all bad or unpronounced phonemes from the sentence and replace them with correct or good phonemes by the normal speaker. The grafting unit is performed according to the following steps proposed by Mohammed Sidi Youcoub *el at*[1,2]

Extract the left phoneme of the bad unit from dysarthric speaker.
Extract the grafted phonemes of the good unit from normal speaker
Cut the right phonemes of the bad unit from dysarthric speaker.
Concatenate and smooth parts above obtained in the three first steps
Lower the amplitude of signal obtained in step 2 and repeat step 4 till to have a good listening

Apart from these conventional methods, there are some other methods which also contribute in supporting the communication aid for dysarthric speakers. Frank Rudzicz [4] has explained TORGOMorph system of acoustic transformation to improve the intelligibility of dysarthric speech.

E. TORGOMorph System

It describes modification to acoustic speech signals produced by speakers with dysarthria in order to make those utterances more intelligible to typical listeners. These modifications include the correction of tempo, the adjustment of format frequencies in sonorants, the removal of aberrant voicing, the deletion of phoneme insertion errors and the replacement of erroneously dropped phonemes. In TORGOMorph system, the components allow for a cascade of one transformation followed by another. The working of these components is explained as follows:

1) *High Pass Filter On Unvoiced Consonants:* The first acoustic modification is based on the observation that unvoiced consonants are improperly voice in in dysarthric speech. In order to correct these mispronunciations the voice bar is filtered out of all acoustic sub sequences annotated as unvoiced consonants

Volume 3 Issue VII, July 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

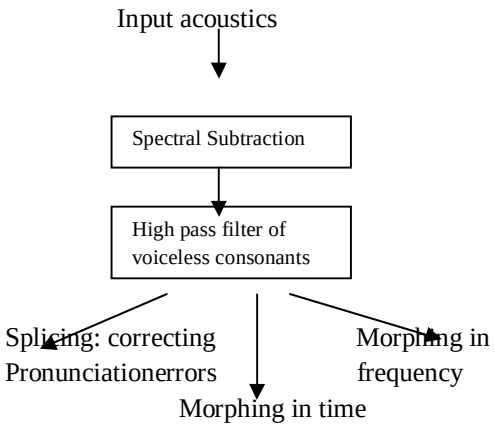


Fig 3 TORGOMorph System

2) *Splicing:* correcting dropped and inserted phoneme errors- When an insertion error is identified the entire associated segment of the signal is simply removed. In the case that the associated segment is not surrounded by silence, adjacent phonemes can be merged together with time-domain pitch-synchronous overlapped. When the deletion of a phoneme is recognized, the associated segment from the aligned synthesized speech is extracted and inserted into the appropriate spot in the dysarthric speech. For all unvoiced fricatives no further action is required.
3) *Morphing In Time*: The vowels uttered by dysarthric speakers are significantly slower than those uttered by typical speakers. In fact, sonorants can be twice as long in dysarthric speech, on average. In this modification, phoneme sequences identified as sonorant are simply contracted in time in order to be equal in extent to the greater of half their original length or the equivalent synthetic phoneme's length. In all cases this involved shortening the dysarthric source sonorant.

The training data required for conventional HMM based ASR is high, especially for speakers who takes great effort in speaking. Jort F. Gemmeke *el at*[3] evaluated an alternative approach which works by mining utterance based representation for recurrent acoustic pattern. This speaker dependent approach developed in ALADIN project, maps these acoustic patterns directly to commands, which means it is language independent and does not require a pre defined vocabulary, grammar or even knowledge of word order in the training data. The ALADIN system has been shown to yield relatively high recognition accuracies even after a single training sample of each word or command.

F. ALADIN

The ALADIN approach works by determining recurrent acoustic patterns in spoken commands, and is based on non negative matrix factorisation approach. NMF is a technique which decomposes a non-negative matrix into the product of two non-negative low-rank matrices. First, the spoken command is converted into an utterance-based vector representation, the acoustic representation. In short, this representation is constructed for each utterance by making a histogram of the co-occurrences of Gaussian posteriors over time, with the Gaussian acoustic model obtained in advance. The acoustic model is estimated through unsupervised k means clustering of the training data, followed by estimating a single full co-variance Gaussian on each cluster.

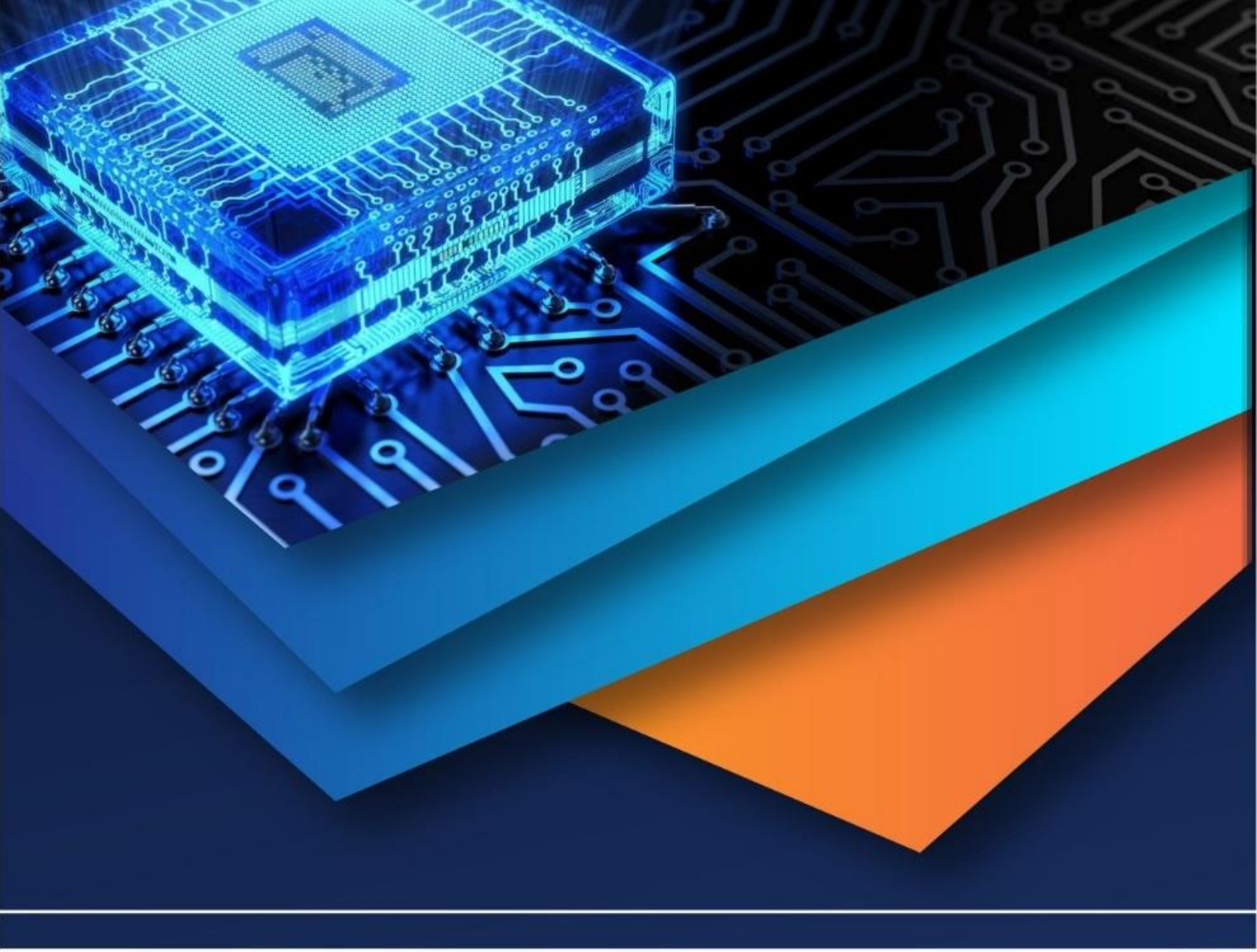
The collection of spoken training commands is concatenated into a matrix, the leftmost matrix in Fig 1, which is then factorised by NMF into a matrix representing recurrent acoustic patterns (the dictionary), and a matrix of activations of these patterns over the training utterances. As visualised by the top half of the matrices, this factorisation is guided (regularized) by the label vectors to ensure that the obtained acoustic patterns correspond to slot values within semantic frames. In addition to acoustic representations that are trained using supervision, a number of acoustic patterns are trained unsupervised to model acoustic phenomena occurring across sentences. These could include breathing sounds and silence, but also words for which no supervision is available.

III. CONCLUSION AND FUTURE ENHANCEMENTS

In this paper, various Speech Assistive Techniques are discussed which are generally used for mild to normal dysarthric speakers. Speech recognition technique with HMM, GMM or MFCC generally uses supervised learning in which the trained data helps in identifying the accurate words which are grafted or concatenated with the original phonemes of the dysarthric speaker. ALADIN approach is used to mine recurrent acoustic patterns from weakly supervised dysarthric speech data, to achieve two goals: 1) Achieving usable recognition accuracies with less training data, in order to minimize the initial effort of the target user, and 2) Achieving usable recognition accuracies with less detailed annotation - training a vocal interface using an unordered list of semantic concepts that are contained in the sentences, rather than a word-by-word transcript. Generally for severe dysarthric speakers these techniques fail to give accuracy more than 90%. The future enhancements can be the speech assistive technology which may use an unsupervised approach and give a better accuracy with minimum time.

REFERENCES

[1] Mohammed Sidi yakoub, Sid-Ahmed Selouani, Douglas O'Shaughnessy "Improving Dysarthric Speech" Intelligibility Through Re-synthesized and Grafted units", May 5-7 2008 Niagara Falls. Canada@ 2008 IEEE
[2] Mohammed Sidi yakoub, Sid-Ahmed Selouani, Douglas O'Shaughnessy "Speech Assistive Technology to Improve the Interaction of Dysarthric Speakers with Machines"March 12-14 @2008 IEEE
[3] "fort F. Gemmeke, Siddharth SehgaP, Stuart Cunningham, Hugo Van hamme Dysarthric Vocal Interfaces with Minimal Training Data" @2014,IEEE
[4] Frank Rudzicz "Acoustic Transformation to Improve the Intelligibility of Dysarthric Speech", Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies, Scotland, UK, July 30, 2011@ 2011 Association for Computational Linguistics
[5] Michelle Cutajar, Edward Gatt, Ivan Grech, Owen Casha, Joseph Micallef "Comparative Study of Adaptive Automatic Recognition of Disordered Speech", IET Signal Process., 2013, Vol. 7
[6] Ahmad A M Abusharaiah, Teddy S Gunawan, Othman O. Khalifa, Mohammad A.M Abusharaiah "English Digits Speech Recognition System based on Hidden Markov Models" @2010 IEEE
[7] Susanne Wagner "Intralingual speech to text conversion in real time: Challenges and Opportunities"EU high level scientific conference series@ 2005 MuTra
[8] D Sasirekha, E Chandra "Text to Speech: A Simple tutorial" @March 2012 IJSCE



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY