ISOLATED WORD RECOGNITION USING NEURAL NETWORK FOR DISORDERED SPEECH

Ankita N. Chadha¹, Mukesh A. Zaveri², Jignesh N. Sarvaiya³
¹,³Department of Electronics, ²Department of Computer Science, Sardar Vallabhbhai National Institute of Technology, Surat, India
¹ankitaism@gmail.com, ²mazaveri@coed.svnit.ac.in, ³jns@eced.svnit.ac.in

ABSTRACT
The society is fully aware of the needs of impaired people. One such impairment resulting due to articulatory or congenital conditions in human speech production mechanism produces a disordered speech. This disordered speech is difficult to interpret and leads to miscommunication. Thus, the speech enabled helping aid is a good alternative to existing devices designed to assist the impaired speakers. Taking this into consideration, this work proposes a spoken word recognition system using a General Regression Neural Network (GRNN) for disordered speech. The complex cepstrum based feature extraction is employed due to their ability to represent the speech production model with mixed phase response. The proposed system is compared with the state of the art techniques. The results are verified using objective and subjective measures. The results suggest that the GRNN based parallel training recognition system performs better in terms of accuracy in contrast to baseline recognition algorithms.

KEY WORDS
Complex Cepstrum, Classification, Disordered Speech Recognition, Regression

1 Introduction
Automatic Speech Recognition (ASR) systems have been an effective research field over many decades. Despite their effectiveness it is yet challenging to recognize the spoken words by a disordered speaker. The disordered speech relates to vocal pathology where speaker suffers from impairment due to failure of control over articulatory organs [1]. It may be due to post surgery effects after operating laryngeal cancer, paralysis, dysarthria, aphasia, spasmodic dysphonia or even cerebral palsy and multiple sclerosis [2]. Also level of impairment has an inverse relation with the accuracy of the recognition system. Hence, designing the framework that assists the impaired individuals for not only reproducing speech but also predicting their performance of improvement are also amongst the scope this research. The applications of disordered speech recognition are voice driven appliances control, interactive interfaces, disorder detection, voice therapist and assessment.

There are very few approaches reported in the literature dedicated to disorder detection and limited data recognition systems with inconsistent results [3, 4]. There is a need to develop systems that adapt according to the speaker and not according to intelligibility of speech.

A generalized speech recognition system involves pre-processing of disordered speech, followed by feature representation methods to reduce data redundancy and computation. Further, an acoustic model comprising of machine learning algorithms are used to complete the offline training process. In testing stage, the test feature vector is applied to the trained model, in order to evaluate the recognition system and obtain the word class [5].

2 State of the Art Techniques
Speech contains information at linguistic and acoustic levels; out of which the acoustical information is associated with the production mechanism of speech. Thus, analysis at the acoustical level produces more reliable outcomes even when there are fine changes in manner of speech over a time span [6]. Based on type of application for disordered speech, the features selection procedure changes, such as for pathology detection, temporal and spectral features are sufficient. The temporal features include those which vary over frame instances, like the average mean, standard deviation, energy and zero crossing rate. While derivatives (deltas), moments, roll-off, decrease and formants belong to the spectral domain [7]. The Linear predictive Coefficients (LPC) based cepstral features are also used for distinguishing between pathological voices [8]. When building systems for recognizing words from disordered speech, Cepstrum and Mel Frequency Cepstral Coefficients (MFCC) are quite popular [9]. Weigl et.al has used the Human Factor Cepstral Coefficients (HFCC) which are based ERB filter bank [10]. The Perceptual Linear Predictive (PLP) along with LPC has also been used for disordered speech recognition [11]. Furthermore, classifiers employed for discriminating disordered speech and recognizing words are Hidden Markov Model (HMM)(both continuous and discrete) [9, 11, 12], Gaussian Mixture Models (GMM),
3 Proposed Speech Recognition System

A generalized automatic speech recognition system is a two-fold process comprising of training and testing stage [15]. Due to the nature of distorted speech, it is assumed that the input words are isolated ones. The training stage has three phases, pre-processing, feature extraction and learning rule. The known utterances are first pre-processed and feature extracted using complex cepstrum based features. The GRNN based learning algorithm is proposed for recognition. The acoustic model formed in training stage is tested using samples which are not employed during training in order to recognize the correct word. This happens to be the testing or recognition stage. Figure 1 shows the diagrammatic representation of our proposed ASR.

3.1 Pre-processing

The first phase of training stage is pre-processing that removes the noise and distortions present in the raw disordered speech sample [9]. The pre-processing begins with filtering the low frequency noise, followed by normalization for eliminating the remaining reverberations such as echoes, dropout, microphone noises and lastly noise due speaking style and physiological structure of speaker. Further, the speech samples are segmented and framed into 30ms windows as they are found to be stationary from 10-30 ms of duration. A 50% overlapping of frames is considered to avoid loss at frame boundaries using hamming window.

3.2 Feature Representation

The pre-processed speech frames are further applied to feature extraction phase. As discussed in the literature, numerous features have been used for disordered speech recognition. Amongst all these features, the complex cepstrum is a good alternative due to the fact that it resembles the human speech production model. The traditional features like LPC represent spectral peaks and fail to symbolize the valleys, forming a minimum phase system [17,18]. While the speech production model is entirely mixed phase, the complex cepstrum based features comprises of minimum phase by the vocal tract and maximum phase by the glottal excitation [19]. This combined representation contributes to a mixed phase system. The complex cepstrum co-efficients are computed by taking Inverse Fourier Transform of the logarithm of Fourier coefficients of speech frame, which can be given as

\[ cc(k) = \mathcal{IF} (\log (F(d_n))) \]  

where \( d_n \) is \( n^{th} \) disordered speech frame, \( cc(k) \) is cepstrum co-efficient and \( k \) is the queue frequency (cepstrum).

3.3 ANN based classification

The last phase of training is completed by employing a learning rule that forms a relationship between the speech patterns for better recognition. There are various recognition techniques used such as GMM, HMM, Neural networks and k-means. The artificial neural networks are gaining importance due to their ability capture non-linearities in the distorted speech [14,20]. In this work, the MLP based recognition used in [14] is considered as a baseline model along with PNN, which has a similar parallel structure like the GRNN but differs in functional behavior [21]. All these neural networks are briefly described in below sections.

3.3.1 Multi-layer Perceptron for Recognition

The MLP neural networks are considered as universal approximators which capture the individuality of speech quite well. There are still issues of slow convergence and local minimas’ that hinder the performance of a MLP classifier [22,23,24]. The MLP based neural network consists of \( n \)-number of hidden layers based on one’s application. Through computations, we found three layered symmetric network to be sufficient for approximating the highly correlated disordered speech samples. The number of neurons \( N \) in the first layer can be calculated as,

\[ N = N_i \times N_o \times \frac{2}{3} \]  

where, \( N_i \) is the size of input and \( N_o \) is size of output.

Similarly, we consider the second hidden layer neurons to be half of first and third layer [25]. The various training algorithms are experimented, out of which the
Scaled Conjugate Gradient (SCG) algorithm is chosen because with increasing number of samples, the size of network increases. Also, this training function has a lower memory requirement and is typically faster than standard gradient descent algorithm \[26\]. Furthermore, these training functions are meant for larger network size. The mapping function, \( F(cc_k) \) of the input feature vector, \( cc_k \) and mean square error, \( err \) are given as

\[
F(cc_k) = \tilde{f} \left( w^i f_{i-1} (w^{i-1} f_{i-2} \ldots) \right) \tag{3}
\]

\[
err = \sum_{i=1}^{C} ||T - F(cc_k)|| \tag{4}
\]

where, \( w^i, w^{i-1}, w^{i-2} \ldots \) represents the weight matrix, \( i \) is number of layers, \( T \) is target class, \( C \) is number of input features. The function \( f_i \) is chosen to be tan-sigmoid. The number of neurons in first, second and third hidden layer are calculated as 65, 30 and 65. Further learning rate of 0.1 is optimized at error value of \( 10^{-4} \).

### 3.3.2 Probabilistic Neural Network for recognition

Another type of single pass learning algorithm is PNN that works on the Bayes rule which accurately represents non-linear decision boundaries. This network is proved to be thousands of times faster than back propagation networks \[27\]. The probability density function (pdf) of each sample is computed and unknown sample’s class is found out accordingly. Their pdf is calculated as

\[
h(y) = \frac{1}{k\sigma} \sum_{i=1}^{k} \left( \frac{y - y_i}{\sigma} \right) \tag{5}
\]

where \( k \) is number of samples, \( \sigma \) is the smoothing factor, \( y \) unknown input and \( y_i \) is \( i^{th} \) input sample.

### 3.3.3 General Regression Neural Network for recognition

The disordered speech data is highly correlated in nature and needs a generalized approximation algorithm that is computationally effective. The GRNN classifier is termed as generalized regression based network with memory that carries out estimates for targets which are continuous in nature. It has a similar structure to PNN, except that the pattern and decision layers are different. It is a one pass learning algorithm with capability of sparse learning \[28\]. The prediction vector \( P(y) \) is given as

\[
P(y) = \frac{\sum_{i=1}^{k} w_i e - g(y, y_i)}{\sum_{i=1}^{k} e - g(y, y_i)} \tag{6}
\]

where \( g(y, y_i) = \sum_{m=1}^{m} \left( \frac{y - y_i}{\sigma} \right)^2 \), \( w_i \) is weight between pattern layer and summation neuron. The value of spread was varied from 0.1 to 3 and \( spread = 1 \) learns the disordered pattern quite well. The GRNN based networks do not need retraining of network as against MLP. Also their training time is smaller than MLP based classifiers with more accurate results and does not get affected by wild points \[29\]. Looking at nature of input feature set it can be concluded that the data is highly correlated and although digitized its dependency on previous and forthcoming values is clearly visible. Hence, it is partially continuous in nature. Moreover, though PNN and GRNN have similar architectures, the GRNN learns the mapping function at a lower spread than PNN.

### 4 Results and discussions

In order to assess our proposed system, we have used the open source Audiovisual UASpeech database which is
developed by University of Illinois containing 15 Speakers with dysarthria and with variations in intelligibility of dysarthric speech from 2% to nearly 95% intelligible [30]. The training and testing data is divided in 90% and 10% respectively. In total 960 Samples are considered for training and around 100 samples for testing the algorithms. The proposed system is evaluated by computing objective measures such as word accuracy rate and confusion scores which validates the classifiers performance. Along with objective measures, the perceptual based subjective test is also conducted in order to compare the computerized scores against the human listeners’ score. The obtained scores of GRNN based system are well rated in comparison to the state of the art MLP and PNN systems.

4.1 Word accuracy

This is the most popular evaluation metric in the automatic speech recognition literature. The number of words correctly classified represents the word accuracy rate [9]. The output predicted class of the recognizer is compared against the target class value for each sample of each word. The overall accuracy for all the words and their respective recognizers are shown in Table 1.

4.2 Confusion Measure

Along with word accuracy, the predicted output class and the required target outputs can be mapped for similarity by plotting a confusion graph. Therefore, the confusion matrix is also called an error matrix between words to be recognized [9]. For a more clear representation, a color map varying between darker shades like red to lighter shades like blue are shown with every confusion graph. The y-axis has the predicted class output while x-axis holds the target class labels for ten words. The recognizer that has a good accuracy will have elements in diagonal of the matrix towards red on the color map indicating word paired to itself. This means that the target matches with the predicted class. While the non-diagonal elements should be more towards blue, indicating word paring with other words with minimum recognition rate. The Figure 2 represents the confusion matrix for GRNN based recognition system, with spread = 1 clearly shows the darker shades being highlighted in the diagonal with accuracy more than 70%. Similarly, Figure 3 and 4 show results for PNN with spread = 3 and spread = 5 respectively. Finally, Figure 5 demonstrates MLP based recognizer with more lighter shades in diagonal than darker shades due to higher confusion between recognized words. Thus, it can be rightly said that the confusion rate amongst intra class is higher for the MLP, while the performance for PNN with spread = 5 performs equivalent to the GRNN with spread = 1. The GRNN based recognition has more concentrated red values lying on the diagonal followed by PNN and lastly the MLP based recognition.

4.3 Perceptual Subjective Test

A perceptual test is also conducted to judge whether a human listener can distinguish between words spoken by disabled speaker with similar accuracy of the computerized ones. Nearly 100 samples were presented to controlled
listeners who have no disability in hearing lying in the age group of 20 to 42 years [6]. The samples were played in isolated environment with loudness adjusted as per the comfortability of the speaker using headsets. The list of words is provided to the listeners to pick from the mentioned words. The results match satisfactory to the proposed work as shown in Table 2 for five listeners.

Table 2. Perceptual subjective test for five Listeners

<table>
<thead>
<tr>
<th>Listener</th>
<th>Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listener 1</td>
<td>68 %</td>
</tr>
<tr>
<td>Listener 2</td>
<td>58 %</td>
</tr>
<tr>
<td>Listener 3</td>
<td>73 %</td>
</tr>
<tr>
<td>Listener 4</td>
<td>69 %</td>
</tr>
<tr>
<td>Listener 5</td>
<td>61 %</td>
</tr>
<tr>
<td>Average</td>
<td>65.8 %</td>
</tr>
</tbody>
</table>

5 Conclusion

The disordered speech based recognizers face challenges in determining the correct spoken word through machine learning approaches. This work evaluates three most profoundly used neural networks namely, GRNN, PNN and MLP networks. These networks were optimized for goal, training function and spread factor. An isolated word recognition system was build using complex cepstrum based features and assessed using objective measures namely, word accuracy and confusion score. The perceptual test is also conducted and yields an average classification accuracy of 65.8%. From the results, it can be concluded that the GRNN based speech recognition systems performs convincingly better with and accuracy of 70% with a lower spread factor while similar results are achieved for a higher spread of 5 using PNN. The MLP based recognition systems face the difficulty in precision of network parameters due to distorted nature of the training data. Also the amount of impairment affects the accuracy of the overall system.

Acknowledgement

The authors are grateful to Mark Hasegawa-Johnson and University of Illinois for building and providing access to UASpeech database for dysarthric speech.

References


