Combining Neural Network and Rule-Based Systems for Dysarthria Diagnosis

James Carmichael, Vincent Wan, Phil Green

Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, UK
{j.carmichael, v.wan, p.green}@dcs.shef.ac.uk

Abstract

This study reports on the development of a diagnostic expert system – incorporating a multilayer perceptron (MLP) – designed to identify any sub-type of dysarthria (loss of neuro-muscular control over the articulators) manifested by a patient undergoing a Frenchay Dysarthria Assessment (FDA) evaluation. If sufficient information is provided describing pathological features of the patient’s speech, the rule-based classifier (RBC) can outperform the MLP in terms of rendering a more accurate and consistent diagnosis. The combination MLP/RBC developed during this study realised an overall improvement in diagnostic accuracy of 9.3% (absolute) for a selection of dysarthric cases, representing a substantial improvement over the benchmark system which – unlike the MLP/RBC – cannot directly process acoustic data.

1. Introduction

The production of well-inflected and clearly enunciated speech depends on the ability to exercise a fine-grained physical control over those organs – collectively known as the articulators – which contribute to the process of oral communication. Any disruption to the neuro-muscular mechanisms responsible for manipulating the articulators will usually result in speech which is acoustically malformed and difficult to understand [3] [4]. The term dysarthria is used to describe those conditions which are responsible for such neuro-muscular articulatory malfunction and the resultant abnormal speech. The correct diagnosis of a dysarthric condition, therefore, is dependent upon the separate assessment of a range of articulatory activities in order to detect any deficiencies therein. This study reports on the prototyping of an automated diagnostic system which is based on the Frenchay Dysarthria Assessment (FDA) evaluation protocol [5] and incorporates both rule-based classification (RBC) and multi-layer perceptron (MLP) technology. The FDA consists of 28 speech measurement sub-tests, the responses to which are graded on a nine-point scale, ranging from “0” to “8”, with zero indicating the lowest performance level and eight the highest 1. Upon completion of an FDA test series, a patient’s test scores are formatted as a histogram (Figure 1), the visual profile of which helps to identify the dysarthria sub-type in evidence.

In the case of the Conversation test (test #28), for example, it is impossible to determine solely via an FDA profile if a poor grade – such as a “D” – has been awarded due to an overly rapid speech rate or an abnormally slow rate, even though such contrasting performances would indicate quite different dysarthria sub-types. Although the FDA provides guidelines to facilitate the correct interpretation of ambiguous data, it is still possible that an inexperienced clinician might find these guidelines insufficient for formulating a consistent and objective diagnosis. The automated diagnostic application – known as the computerised FDA (CFDA) [2] – developed during the course of this investigation aims to assist the clinician in producing a more accurate and consistent evaluation of an observed dysarthric condition by using non-intrusive acoustic signal processing techniques to measure articulatory dysfunction. These measurements are then classified using RBC and MLP pattern recognition approaches to identify the specific dysarthria sub-type in evidence.

The CFDA’s success in facilitating more accurate dysarthria diagnosis will be evaluated by comparing it with a previous FDA-based diagnostic software application known as the Frenchay Dysarthria Test (FDT). The following section describes the FDT’s functionality and its accuracy; section 3 details the CFDA’s hybrid NLP-RBC architecture; sections 4 and 5 discuss the experiments devised to demonstrate the CFDA’s improved performance using the FDT as the benchmark standard. Section 5 also concludes the paper with a recommendation for future research activities in this area.

1 The nine intervals on this scale are also expressed as the letter grades “E”, “E+”, “D”, “D+”, “C”, “C+”, “B”, “B+” and “A”, with “E” corresponding to zero and “A” to the highest score.
2. Prior Work in Automated Dysarthria Diagnosis

The FDT, implemented by Roworth (1990), is the first and only attempt (apart from the CFDA) to automate the dysarthria diagnostic process. The FDT application – which uses the linear discriminant analysis (LDA) approach adopted by Enderby [4] – takes as its sole input the patient’s FDA grade profile and returns the dysarthria sub-type which best corresponds to said profile. When tested on FDA diagnostic data from the Enderby Frenchay Assessment (EFA) corpus, the FDT’s overall classification accuracy was found to be 76.82%, although this accuracy was not uniform across the five dysarthria categories [2], as detailed in Table 1.

Table 1: FDT Classification Correctness per Dysarthria Type

<table>
<thead>
<tr>
<th>Dysarthria Sub-type</th>
<th>FDT Classification Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ataxic</td>
<td>42.9</td>
</tr>
<tr>
<td>Extrapyramidal</td>
<td>84.2</td>
</tr>
<tr>
<td>Flaccid</td>
<td>100</td>
</tr>
<tr>
<td>Mixed</td>
<td>78.6</td>
</tr>
<tr>
<td>Spastic</td>
<td>78.4</td>
</tr>
<tr>
<td><strong>Average Accuracy</strong></td>
<td><strong>76.8</strong></td>
</tr>
</tbody>
</table>

This level of accuracy is low considering that the data used to train the FDT’s diagnostic algorithms is the entire EFA corpus itself, which is the same corpus used to generate the test results presented in Table 1. Higher levels of accuracy are usually expected when an automatic classifier’s training and testing sets are identical. It is also perplexing that the FDT only offers a single hypothesis as output for a given FDA grade profile, especially since the paper-based FDA evaluation protocol [3] takes care to remind the clinician that it is not always clear which dysarthria type is responsible for a patient’s symptoms, thus it is advisable to consider more than one candidate.

It is possible that this below-par performance is partially due to the fact that the FDT does not consider the actual quality of any articulatory behaviour (such as the speech rate phenomenon described in the previous section) which inspires the grades composing a patient’s FDA profile. Such inability to process all but one type of data may be one of the principal reasons for the FDT’s high incidence of classification inaccuracy for cases of Ataxic dysarthria documented in the EFA data corpus. Indeed, the FDT’s procedure for computing a diagnosis using only the patient’s grade profile is not commensurate with the paper-based FDA’s recommendation that the manifestation of certain articulatory dysfunctions – e.g. breathy voice – should be accorded special significance since they are closely associated with particular dysarthria sub-types. In an attempt to remedy the FDT’s shortcomings, the CFDA’s diagnostic algorithms attempt to process any accessible qualitative and quantitative diagnostic information not explicitly represented by the FDA profile grades. The

FDAs multi-faceted analytical approach to dysarthria diagnosis (using speech measurement techniques along with RBC and MLP analysis) will now be considered in more detail.

3. CFDA Multi-Faceted Diagnosis

Unlike the FDT which is only an LDA classifier, the CFDA can directly capture and evaluate a patient’s speech data since it is a complete diagnostic application meant to replace the paper-based FDA. To this end, the CFDA incorporates digital audio signal analysis techniques which measure various articulatory functions such as voice quality and the ability to vary pitch. The severity of any speech-related pathological phenomena can thus be quantitatively described and these descriptions are then processed by the CFDA’s rule-based classifier to help identify the dysarthria sub-type causing the symptoms. When assessing speech rate, for example, the CFDA can be used to record a sample of the patient’s speech and – via the use of the mrate algorithm [6] – measure the rate thereof in terms of syllables per second (sps). The CFDA can also convert this sps measurement into a conventional FDA grade (e.g. “A” or “8”) for the purposes of creating a grade profile. The CFDA’s capacity to generate and process multi-dimensional data input therefore reduces the possibility of an incorrect diagnosis due to misinterpretation of the test results. The automatic methods used to classify these multi-faceted test results are discussed in the sections immediately following.

3.1. The MLP component of CFDA diagnosis

The CFDA’s MLP serves the same purpose as the FDT’s LDA: both take a 28-point FDA grade profile as their input in order to compute a diagnostic hypothesis. Since supplementary information (e.g. medical notes made by the clinician administering the test) is not always available, it is not practical to include such data into for MLP processing. For the EFA corpus, however, it was demonstrated that MLP analysis proved superior to that of LDA classifiers when tested on unseen data (i.e. data not included in the classifier’s training set). The CFDA’s neural network architecture is as follows:

- The network itself is a 3-layer MLP with tanh activation functions at the hidden layer and softmax at the output that estimates the posterior probabilities of the five dysarthria types.
- Due to data scarcity (the EFA corpus of 85 cases being the only available source), the MLP was trained and tested using the jack-knifing approach whereby the system is trained on all but one of the data points and that excluded data point then constitutes the testing set. By training a new MLP each time all 85 can be used as legitimate unseen test cases.

Of course, it is possible to over-optimise an MLP’s parameters so that it will achieve a 100% classification accuracy on the training data but such overtraining normally results in poor generalisation. To help avoid overtraining, the number of parameters in the MLP was reduced to be fewer than those in the training data. To this end, principal components analysis (PCA) was first applied to reduce the EFA data dimensionality.
The MLP is trained in NETLAB [1] [7] using conjugate gradient descent with weight decay \( W \) to help prevent overtraining. Many possible values of \( I, H \) and \( W \) were tested and the best configuration was found to be \( I=5, H=5 \) and \( W=10^{-6} \) leading to an MLP with 60 weights. Figure 2 shows the accuracy of the MLP averaged over 10 independent training runs, each initialised with random weights. The dotted lines indicate ±1 standard deviation. The figure indicates that an average accuracy of 51.8% (with a maximum accuracy of 60.2%) can be achieved after approximately 20 iterations. In some cases, it is possible – as demonstrated in the experimental results presented in section 5 – to better this performance via the input of the rule-based classifier. A more thorough analysis of the rule-based classifier, both in stand-alone mode and in conjunction with the MLP, now merits closer attention.

3.2. RBC multi-faceted data processing

The CFDA’s rule-based classifier is, essentially, a series of manually constructed decision trees (such as the one depicted in Figure 3) that encode the diagnostic knowledge of a group of expert clinicians\(^3\), figuring among whom is the author of the paper-based FDA; the RBC is therefore designed to emulate the decision-making behaviour of an expert when analysing a patient’s performance during an FDA evaluation. The RBC is capable of processing multi-dimensional input by gathering information directly from several sources, including:

(i) medical notes made by the presiding clinician while observing the patient under test conditions; these notes are recorded in a structured electronic format using the CFDA application itself;

(ii) answers to pre-formulated questions directly posed to the clinician administering the FDA evaluation. These questions serve to elicit further details about certain observed pathological behaviours;

(iii) measurements of spectral features extracted from audio recordings of the patient’s oral responses to FDA stimuli.

Once all accessible information has been gathered from the abovementioned sources, the RBC channels such multifaceted data through its decision trees to generate a diagnosis. In the event of data inconsistencies producing conflicting outputs from different decision trees (e.g. the diagnostic hypothesis from the speech-rate decision tree is totally at odds with the hypothesis from the voice-quality decision tree), the CFDA will warn of such discrepancies and adopt one of two courses of action: (i) further questions will be posed to the examining clinician in order to clarify any ambiguities or (ii) if such clarification is not possible, then the CFDA will advise that a reliable diagnosis cannot be extracted from the available information.

![Example of an RBC Decision Tree](image)

In terms of its performance when tested solely on the EFA corpus’ FDA grade profile data, the RBC’s overall diagnostic correctness was 54.1%. Unfortunately it was not possible to apply the RBC’s multifaceted data processing capability to every case within the EFA corpus since the medical notes which accompany the grade profiles are not always comprehensive; in fact, there are only ten cases with notes containing sufficient information to entirely satisfy the RBC’s information requirements (i.e. to provide answers to all of the questions posed by the classifier). The fact that the RBC correctly diagnosed all of these ten cases confirms the supposition that this classifier does not perform optimally when restricted to only one type of data. In an effort to obtain a better performance under conditions of multi-dimensional data scarcity, this study reports on the development of a hybrid diagnostic system combining the advantages of the MLP and RBC so as to offer state-of-the-art performance under diverse situations.

3.2.1. Designing a hybrid MLP-RBC diagnostic classifier

As demonstrated previously, the MLP can achieve superior accuracy if the only diagnostic information available is a 28-point grade profile, while the RBC proves the better classifier when provided with an ample supply of supplementary data. Of course, it is often the case that an FDA interview will yield information which is neither absolutely minimal nor adequate in every aspect for the RBC’s purposes. It would be useful, therefore, to devise a diagnostic system tolerant of data scarcity and yet able to fully exploit all additional information sources. In order to achieve this objective, a protocol has been devised whereby the classification hypotheses for both the MLP and RBC systems are assigned different weightings dependent on the quantity and quality of available diagnostic data. If, for example, the CFDA measures a patient’s speech rate as being particularly rapid (i.e. more than 4 syllables per second), this indicator will strongly bias the outcome of the CFDA’s final diagnostic hypothesis and may even negate a

\(^3\) This expert knowledge was captured and documented in a series of interviews with the clinicians in question.
conflicting hypothesis offered by the MLP. The calculation of such weightings/biases will now be more closely scrutinised.

4. Operating the 2 Classifiers in Conjunction

Of the 85 FDA profiles in the EFA corpus, 40 featured accompanying medical notes with at least the minimal amount of information (i.e. including at least one symptom description which can be parsed by the RBC decision trees) to allow meaningful processing by the RBC. These 40 cases were used to test a hybrid RBC-MLP system observing the following protocol in the event of a discrepancy between the outputs of the two classifiers: (i) the RBC’s output is preferred if the medical notes confirm the presence of a symptom strongly associated with a specific dysarthria sub-type – e.g., uncontrollably rapid speech would be a high-probability indicator of hypokinetic dysarthria; (ii) in the absence of any such definitive symptoms, the top three results for each of the two classifiers are compared to determine the extent of discrepancy and then the two hypotheses most consistent with the symptoms reported in the medical notes are selected as the final candidates. To cite an example; for one of the 40 cases selected, the MLP’s hypotheses – in decreasing order of likelihood – were (1) Mixed dysarthria (2) Spastic (3) Extrapyramidal, while the RBC’s top three candidates were (1) Mixed dysarthria (2) Flaccid (3) Ataxic. The accompanying medical notes indicated that the patient’s speech was mildly hypernasal but his face was markedly asymmetric, a symptom combination most consistent with mixed dysarthria but less so with the purely spastic variety. The system therefore decides – via reference to a look-up table linking symptom severity with specific disease sub-types – that the two most likely dysarthria types responsible for the patient’s condition are (1) mixed and (2) spastic, and these two choices are offered as the final output. The hybrid system’s two other hypotheses – i.e., Extrapyramidal and Ataxic – are discarded since they are the least consistent with the reported symptoms. It will be noted that this hybrid system suggests two diagnostic hypotheses rather than one as is the case with the FDT. In the context of dysarthria diagnosis, this dual hypothesis output does not represent a conflict since it is not always possible to distinguish between two closely related disease subtypes solely by examining the symptoms. Table 2 compares the diagnostic output accuracy of the FDT and MLP/RBC systems for the abovementioned 40 cases.

Table 2: MLP/RBC vs. FDT Classifier Accuracy for 40 Selected Cases from the EFA Corpus

<table>
<thead>
<tr>
<th>Dysarthria Sub-type</th>
<th>FDT Classification Accuracy (%)</th>
<th>RBC/MLP Classification Accuracy (%)</th>
<th>RBC/MLP Classification Accuracy (%) 1st or 2nd hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ataxic</td>
<td>53.1</td>
<td>75.8</td>
<td>88.4</td>
</tr>
<tr>
<td>Extrapyramidal</td>
<td>78.7</td>
<td>86.0</td>
<td>96.2</td>
</tr>
<tr>
<td>Flaccid</td>
<td>94.2</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Mixed</td>
<td>82.2</td>
<td>78.0</td>
<td>84.6</td>
</tr>
<tr>
<td>Spastic</td>
<td>70.8</td>
<td>85.6</td>
<td>94.6</td>
</tr>
<tr>
<td>Avg. Accuracy</td>
<td>75.8</td>
<td>85.1</td>
<td>92.8</td>
</tr>
</tbody>
</table>

5. Conclusions and Future Work

The RBC/MLP classifier’s 9.3% overall improvement in accuracy over the FDT for the first-choice hypothesis is almost doubled if both the first and second choices are considered. Furthermore, it must be noted that the level of supplementary documentation for the majority of the selected 40 cases barely met the minimum data processing requirements for the RBC. This data scarcity is partially due to the fact that the acoustic recordings of the patients’ performances when responding to the FDA test stimuli have not been preserved, a situation which would have been avoided if the diagnostic interviews had been conducted using the computerised FDA which – as mentioned in section 1 – incorporates acoustic signal processing techniques to measure speech-related pathological features. In various experiments comparing the grading of dysarthric speech samples by expert clinicians and CFDA automated analytical procedures, the level of machine-human correlation exceeded 0.8 for seven of the FDA’s twelve assessment procedures (e.g. assessment of speech rate) which are based primarily on auditory evaluation. Unfortunately, these speech samples used to demonstrate the validity of CFDA objective measurements originated from individuals whose dysarthric conditions have not – at the time of writing – been formally confirmed by an independent third party. When such confirmation has been obtained, the detailed results of these studies will be published for wider scrutiny. It is also envisaged that newly gathered data – the quantity of which will be increased by the further recruitment of more patients – will be added to the EFA corpus and provide clinicians with much needed resources to enhance their professional development in this particular sub-discipline of speech therapy.

6. Acknowledgements

This study, dedicated to the memory of Gladys Martineau Carmichael, would not have been possible without the financial support of the University of Sheffield’s School of Health and Related Research (ScHARR).

7. References