Revisiting Dysarthria Assessment Intelligibility Metrics

James Carmichael, Phil Green

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

Abstract

This study reports on the development of an automated isolated-word intelligibility metric system designed to improve the scoring consistency and reliability of the Frenchay Dysarthria Assessment Test (FDA). The proposed intelligibility measurements are based on the probabilistic likelihood scores derived from the forced alignment of the dysarthric speech to whole-word hidden Markov models (HMMs) trained on data from a variety of normal speakers. The hypothesis is that these probability scores are correlated to the decoding effort made by naïve listeners when trying to comprehend dysarthric utterances. Initial results indicate that the scores returned from these composite measurements provide a more fine-grained assessment of a given dysarthric individual’s oral communicative competence when compared with traditional ‘right-or-wrong’ scoring of expert listeners’ transcriptions of dysarthric speech samples.

1. Introduction

As is normally the case with evaluation based on subjective criteria, robust and consistent diagnosis of speech disorders is frequently hampered by a lack of consensus among experts vis-à-vis the interpretation of assessment data [3]. Inconsistent classification / reclassification by a given evaluator for the same data set is not uncommon and can vary in excess of 30%, particularly in the case of intelligibility assessments [1] [3]. This study reports on the implementation of a computer-based objective metric system for the evaluation of dysarthria – a condition defined as any motor-neuron based loss of control of the articulators resulting in abnormal speech production. This particular computerised system, although designed specifically for the Frenchay Dysarthria Assessment Test (FDA) isolated-word test [2], could be modified for use with any intelligibility test which uses a closed set of vocabulary items.

In a typical test scenario, the speaker being evaluated pronounces a series of words and sentences which a panel of judges then attempts to interpret. The speaker’s intelligibility is computed as a percentage representing the average number of correct interpretations. The disadvantage of such an assessment system is its inherent psychometric weakness: a listener’s capacity to understand a given speaker is influenced by several factors including the level of previous exposure to said speaker’s accent, cultural background and personal association with the individual being assessed. Carmichael and Green have also demonstrated that there is significant discrepancy in speech perception between expert listeners (those with formal training in a relevant discipline, such as linguistics) and non-experts [1]. In the particular case of the FDA intelligibility tests, psychometric inconsistency is further compounded by certain anomalies in the evaluation criteria which instruct the judge to award:

- an ‘A’ grade if all 10 utterances in the test set are interpreted correctly and easily intelligible.
- a ‘B’ if all 10 utterances are correctly interpreted but particular care in listening is required
- a ‘C’ if 6-9 words/phrases are understood.
- a ‘D’ if 3-5 utterances are correctly interpreted.
- an ‘E’ if less than 3 words are understood.

The implicit assumption here is that if the judge fails to decipher all ten of the speaker’s utterances, it automatically follows that particular care was needed to interpret any of the ones that were actually understood; this assumption is not always valid. The proposed automatic speech recognition (ASR)-based metric system discussed in this study seeks to provide the tools for a more fine-grained analysis of intelligibility results by using forced-alignment likelihood scores, the function of which are discussed in the following section.

2. Intelligibility Testing via HMM-based ASR

In relation to intelligibility testing, one of the principal disadvantages of non-adaptive speaker-independent ASR becomes an invaluable tool for gauging the extent to which a speaker can be understood by the naïve listener (someone with no previous exposure to that particular individual’s speech style). To achieve the best possible recognition rate for a given population, hidden Markov model (HMM)-based ASR requires a cross-section of acoustic training data which best typifies the speech of the target group. In effect, the recogniser creates ‘everyman’ archetypal phone and/or word models that embody the statistical estimates of the training data acoustics. Any incoming utterance is compared with these models to determine which model is most likely to have generated the ‘unknown’ speech signal. If it is known beforehand what will be said, one may pre-select the model to which the utterance corresponds in order to gauge how well the utterance’s acoustic profile fits that of the intended model. This process, known as forced alignment, produces a goodness of fit (GOF) likelihood, typically expressed as a logarithmic value, $10^{-x}$ (10 being the maximum). The smaller the GOF, the higher the variance from the norm. The likelihood of any specific individual producing speech aligning well with the recogniser’s acoustic models declines in direct correlation to that individual’s speaking style variance from the large-group statistical norm. We hypothesise, therefore, that these GOF ratings will be systematically related to speaker intelligibility.

This variance-from-norm may be considered analogous to the decoding effort involved in oral communication: understanding speech, of course, is not simply a binary right-or-wrong choice. Even when correctly interpreted, some speaking styles demand greater effort to correctly decipher than others for any given listener. The reality of different speaking styles demanding varying degrees of decoding effort...
implies that intelligibility incorporates the notion of the ideal utterance: every listener has some concept of what a word or series of words should ‘ideally’ sound like and the greater a speaker’s variation from this expected acoustic production, the greater the decoding effort required.

In what would appear to be the first attempt to model this decoding effort using objective measures, Carmichael and Green, trained isolated-word intelligibility estimators (i.e. recognisers operating in forced alignment mode) with data from a variety of normal speakers using the following speech processing configurations [6]:

- Whole word (as opposed to phone level) modelling.
- Acoustic vectors comprising 12 Mel frequency cepstral coefficients (MFCCs) plus their delta values.
- 11 HMM states per word model with three Gaussian mixtures per HMM state.

Dysarthric speech samples processed by these estimators returned forced alignment log likelihoods on average 25 times lower than those for normal speech samples. This disparity was also reflected in the subjective assessments of a panel of naïve and expert listeners assigning comprehension difficulty ratings to the same speech samples. Due to a lack of appropriate dysarthric speech test data at the time, this initial proof-of-concept study, although encouraging, could not address certain key issues, namely the intelligibility estimator’s capacity to differentiate between normal speech and milder forms of dysarthria. (The speech used in the 2003 study all came from individuals suffering from severe forms of the condition). Additionally, the speech samples used in this preliminary study did not include most of the 50 items in the FDA isolated word intelligibility test set due to lack of appropriate dysarthric speech data at the time.

This current study seeks to redress the shortcomings of its predecessor by evaluating the speech of a wider range of dysarthria types and levels of severity. Given the evidence that HMM recognisers analyse speech data in ways substantially different to human perceptual decoding [5], a pseudo-phone level intelligibility estimator was built in an attempt to isolate specific short duration phonemes (such as plosives) that provide critical recognition cues. The following section elaborates on the particulars of experimental design.

3. Experimental Design

If one accepts the hypothesis that intelligibility is synonymous with decoding effort and is dependent upon the level of familiarity with the speaking style being assessed, the rationale for using HMM-based ASR to model said decoding effort rests on the premise that – if sufficiently varied and representative of the target population’s speech patterns – a recogniser’s training data will produce acoustic models which can be considered analogous to a naïve listener’s experience and expectations of what a particular word or word combination should sound like. Speech comprehension, of course, is not symmetrical with speech production: one can normally understand more speaking styles than one can produce. Apart from the ‘native’ speaking style, the speaking style range comprehensible to members of a given community depends on that community’s socio-linguistic interaction with other groups. In this study, we have attempted to model – insofar as the FDA isolated-word intelligibility test is concerned – the speech production and comprehension characteristics of the British South Yorkshire region. The 50-word vocabulary recognisers’ training data consisted of speech samples recorded from a representative demographic and age-group cross-section of the male population1. The test data comprised 30 FDA intelligibility test words recorded by four dysarthric speakers with the following pathological classifications (as classified by experts using FDA analysis):-

- **Dysarthric Speaker 1**: manifests a mixture of spastic and ataxic dysarthria typologies with symptoms being classified as severe, having a subjective intelligibility rating of 10%.
- **Dysarthric Speaker 2**: Cerbellar Ataxic dysarthria with mild speech impairment and intelligibility of 76%.
- **Dysarthric Speakers 3 and 4**: Spastic-type dysarthric condition arising from Cerebral Palsy; their speech production is moderately affected and have scored 65% and 62% respectively in the FDA isolated word subjective tests.

The test data was evaluated by both the automated systems and a panel of listeners with the following range of abilities:

- **Listener 1**: Speech therapist and phonetician currently administering therapy to the dysarthric speakers whose speech samples feature in this study.
- **Listener 2**: Speech therapist and phonetician previously involved in administering therapy to severe dysarthrics but not familiar with the four dysarthrics speakers under review here.
- **Listeners 3 - 8**: Naïve listeners with only casual exposure to dysarthric speech in general and no previous acquaintance with the dysarthric speakers in this study.

In keeping with FDA test conventions, the assessors heard each utterance once only, transcribing what they had heard and assigning a subjective intelligibility assessment grade based on a five point scale with the following criteria (modified for this study):

- **Grade 1**: exceptionally clear speech, such as from a trained broadcast announcer;
- **Grade 2**: normal clarity, typical of everyday conversation;
- **Grade 3**: a moderately greater effort than normal needed to identify the utterance;
- **Grade 4**: barely intelligible, maximum effort needed to decipher what was said;
- **Grade 5**: Unintelligible, the speech sound is not identifiable as a word or series of words known to the listener.

Apart from word-level processing, a pseudo phone-level intelligibility estimator was also built in an attempt to isolate those particular phonemes (often word-initial consonants in the case of the FDA test set) that serve as important cues for subjective assessment [4], examples of such recognition-critical phones being the initial /b/ plosive and /t/ consonant cluster in “Bubble” and “Stubble” respectively. Due to data scarcity, only thirty items of the FDA 50-word set were used. For the purposes of inter-speaker comparison for a given stimulus, more than one recording was made for some of the minimal pairs among the thirty items. The correlation between subjective evaluation and the recogniser’s measurements are presented in the section that follows.

4. Experimental Results

As usually obtains with intelligibility assessments, inter-asser variability was significant with the listener recording the most correct interpretations (28 out of 50 for Expert Listener 1) doubling that of the lowest (14 out of 50 for Naïve Listener 3). In terms of group averages, the expert listeners

---

1 Due to time constraints and the fact that the dysarthric speech used in this study originates exclusively from male speakers, the recognisers built for this experiment were trained on male speech only; the 20 speech samples used per word originated from South Yorkshire speakers comprising of 2 pre-adolescents, 4 adolescents, 10 adults aged between 21 and 49, and four adults ranging in age from 50 to 78.
recorded a 54% recognition rate as contrasted with 37% for the naïve listeners. Table 1 presents individual assessors’ scores relative to the four dysarthric speakers under review:

**Table 1: Average Subjective Evaluations & Response Times**

<table>
<thead>
<tr>
<th>Listener</th>
<th>Speaker 1</th>
<th>Speaker 2</th>
<th>Speaker 3</th>
<th>Speaker 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listener 1</td>
<td>Very hard</td>
<td>Fairly hard</td>
<td>Fairly hard</td>
<td>Fairly hard</td>
</tr>
<tr>
<td>Listener 2</td>
<td>Unintell.</td>
<td>Fairly hard</td>
<td>Very hard</td>
<td>Very hard</td>
</tr>
<tr>
<td>Listener 3</td>
<td>Unintell.</td>
<td>Very hard</td>
<td>Unintell.</td>
<td>Very hard</td>
</tr>
<tr>
<td>Listener 4</td>
<td>Unintell.</td>
<td>Fairly hard</td>
<td>Very hard</td>
<td>Fairly Hard</td>
</tr>
<tr>
<td>Listener 5</td>
<td>Unintell.</td>
<td>Very hard</td>
<td>Very hard</td>
<td>Unintell</td>
</tr>
<tr>
<td>Listener 6</td>
<td>Very hard</td>
<td>Very hard</td>
<td>Very hard</td>
<td>Fairly hard</td>
</tr>
<tr>
<td>Listener 7</td>
<td>Very hard</td>
<td>Very hard</td>
<td>Very hard</td>
<td>Fairly hard</td>
</tr>
<tr>
<td>Listener 8</td>
<td>Fairly hard</td>
<td>Very hard</td>
<td>Very hard</td>
<td>Very hard</td>
</tr>
<tr>
<td>Average Resp. Time</td>
<td>7.1</td>
<td>4.5</td>
<td>5.3</td>
<td>4.7</td>
</tr>
</tbody>
</table>

4.1. HMM-ASR measurements for the Test Corpus

The word-level intelligibility estimator’s behaviour, when compared to the panel of listeners, was inconsistent in certain aspects. In terms of right-wrong recognition decisions for specific utterances, the whole-word estimator’s performance roughly approximated perceptual judgements in most cases: utterances which were correctly understood and adjudged fairly difficult by the majority of naïve listeners (with normalised response times ranging from 3.5 to 5 seconds) usually received GOF scores which varied on average by a factor of ~22 from the maximum likelihood utterance. Utterances perceived as very hard to decode averaged a GOF variance of ~36. Moreover, there were four utterances (originating from the mild dysarthric) deemed to be of normal speech quality and varied by less than -10. Notwithstanding this overall harmonisation of whole-word GOF scoring and subjective assessment, there were instances where the word-level intelligibility estimator’s measurements were at odds with naïve listener assessment. In some cases, the recogniser’s non-speech models proved inadequate filters for some inordinately lengthy pre-speech vocalisations that characterise certain types of dysarthria but which – judging by the number of correct responses – do not defeat the compensation mechanisms of the naïve listener. An example of such pre-speech is presented in Figure 2.

**Figure 2: Speaker 2’s “Jacket” Utterance**

---

2 Listeners 2, 3 and 4 recorded no correct interpretations (i.e. 0%) for Speaker 1, hence the lack of a blue bar chart element in the 2nd, 3rd and 4th histogram clusters

3 This response reaction time was then normalised taking into consideration the judge’s average response times for other utterances of the same perceived difficulty level.

4 For the subjective measurements, the modal average is reported (i.e. the most frequently occurring grade for those utterances which were correctly interpreted).
In this instance, as well as several others, the pre-speech segments so dominated the signal that it caused the speech detection mechanism to misjudge word onset and therefore compute a GOF score indicating severe abnormality, an assessment at variance with perceptual assessments. Table 2 lists a sample of GOF scores for utterances produced by the mild and moderately dysarthrics speakers with the anomalous score highlighted. Devising robust models to filter out dysarthric pre-speech remains a problematic issue in the design of intelligibility estimators. The uneven performance returned by whole-word analysis re the assessment of ambiguous phonemes was in some measure bettered by the pseudo-phone level intelligibility estimator, as discussed in the sub-section that follows.

4.1.1. Pseudo-phone level intelligibility analysis

Based on Greenberg’s findings [4] regarding emphasis patterns on syllable-initial phones in words, sub-word level intelligibility estimators were built for seven items of the test corpus, namely the minimal pairs “Bubble/Double”, “Dark/Park”, “Go/Glow” along with “Warm” and “Here”. Departing from standard phone-level ASR processing where the utterance is analysed as phonemically distinct units, a distinction was made simply between syllable onset – e.g. [gl] of “Glow” – and the rest of the speech signal. The force-alignment log likelihoods returned by each segment were accorded equal weighting in computing the overall GOF rating. For the minimal pair items in question, the pseudo-phone level intelligibility estimators better emulated the behaviour of the naïve listener group – in the context of decoding effort evaluations – than their whole-word counterparts. Syllable-initial segment analysis of the dysarthric utterance representing “Bubble”, for example, returned GOF variance scores of -33 for the first syllable and -27 for the second, resulting in an overall -30 GOF rating by the phone level estimator – as opposed to the whole word intelligibility estimator’s GOF rating of -21 – which correlated better with the naïve listener group average response time of 6.3 seconds and subjective intelligibility rating of “Very hard”. Similarly, several utterances representing “Glow” and “Go” – which were perceived by five out of the six naïve listeners as being unintelligible – were accorded GOF ratings aligning better with perceptual judgements using pseudo-phone level analysis.

5. Conclusions and Future Work

Although the hypothesis presuming a correlation between GOF scores and naïve listener decoding effort appears to be tenable, this research team is still in the process of formulating the actual conversion algorithm to translate these GOF scores into a metric which the targeted users – speech therapists – would find more palatable. Additionally, it is not clear whether there are absolute GOF boundary values demarcating various severity levels of dysarthria (e.g. if a GOF value range can be identified outside of which a speech sample can be definitely considered to be moderately as opposed to mildly dysarthric).

Perhaps of greater import than boundary value demarcation is the effect of prosodic and suprasegmental variation (i.e. accent) on intelligibility: in this study, the normal speech samples used to train the intelligibility estimator came from speakers resident in the UK’s south Yorkshire region. Intelligibility estimators trained on speech samples from different regions in different territories may analyse speech in significantly different ways and this uncertainty can only be resolved by conducting similar studies on a larger scale using a greater diversity of speech training data. The influence of accent variation would, of course, be more pronounced in word sequences, for which there is a separate FDA sentence test. It is this research team’s intention to investigate speaker accent influence on intelligibility in the near future.

6. Acknowledgements

This study would not have been possible without the dysarthric speech samples contributed by the EC Framework 5 OLP programme and previous dysarthria-specific ASR research conducted by the UK National Health Service’s STARDUST project.

7. References


Table 2: Naïve Group Response Times and Intelligibility Estimator’s GOF Scores for Six Selected Utterances

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Average Resp. Time</th>
<th>GOF Variance from Norm</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Jacket” (Speaker 3)</td>
<td>6.0</td>
<td>-20</td>
</tr>
<tr>
<td>“Jacket” (Speaker 2)</td>
<td>4.9</td>
<td>-35</td>
</tr>
<tr>
<td>“Play” (Speaker 2)</td>
<td>4.0</td>
<td>-15</td>
</tr>
<tr>
<td>“Play” (Speaker 4)</td>
<td>5.8</td>
<td>-18</td>
</tr>
<tr>
<td>“Single” (Speaker 2)</td>
<td>3.1</td>
<td>-15</td>
</tr>
<tr>
<td>“Single” (Speaker 3)</td>
<td>4.4</td>
<td>-22</td>
</tr>
</tbody>
</table>