Published in final edited form as:
J Speech Lang Hear Res. 1997 October ; 40(5): 1073–1084.

Development of A Two-Stage Procedure for the Automatic Recognition of Dysfluencies in the Speech of Children Who Stutter: I. Psychometric Procedures Appropriate for Selection of Training Material for Lexical Dysfluency Classifiers

Peter Howell, **Stevie Sackin**, and **Kazan Glenn**
University College London

Abstract

This program of work is intended to develop automatic recognition procedures to locate and assess stuttered dysfluencies. This and the following article together, develop and test recognizers for repetitions and prolongations. The automatic recognizers classify the speech in two stages: In the first, the speech is segmented and in the second the segments are categorized. The units that are segmented are words. Here assessments by human judges on the speech of 12 children who stutter are described using a corresponding procedure. The accuracy of word boundary placement across judges, categorization of the words as fluent, repetition or prolongation, and duration of the different fluency categories are reported. These measures allow reliable instances of repetitions and prolongations to be selected for training and assessing the recognizers in the subsequent paper.

Keywords

stuttering frequency counts; artificial neural networks; psychophysical assessment of stuttered speech

Stuttering is a disorder of speech communication. Traditionally, it is characterized by the occurrence of certain canonical dysfluency types interspersed with episodes of speech that are, at least superficially, fluent. The types of dysfluencies that Johnson et al. (1959) employed are; category a. interjections (extraneous sounds and words such as “uh” and “well”); category b. revisions (the change in content or grammatical structure of a phrase or pronunciation of a word as in “There was once a young dog, no, a young rat named Arthur”); category c. incomplete phrases (the content not completed); category d. phrase-repetitions; category e. word-repetitions; category f. part-word-repetitions; category g. prolonged sounds (sounds judged to be unduly prolonged); category h. broken words (words not completely pronounced).

Conventional ways of making stuttering assessments are to count the occurrences of these types of dysfluencies and express them either as the number of dysfluent words as a proportion of all words in a passage, or measure the time the dysfluencies take compared with the duration of the entire passage (Starkweather, Gottwald, & Halfond, 1990). The main difficulties in making such counts are that (a) they are time consuming to make and (b) there is poor agreement when different judges make counts on the same material (Kully &

Address all correspondence to Peter Howell, Department of Psychology, University College London, Gower Street, London WC1E 6BT, England..

Boberg, 1988). The latter potentially brings claims about what effect a treatment has achieved, whether a child should be diagnosed as stuttering, and so on, into question.

It is vital, then, that assessments are improved. The first difficulty could be circumvented by automating the assessments. However, this presupposes that the problem of low interjudge agreement is sorted out too; there would be little point implementing an automatic version of a procedure that could not be validated due to disagreement between judges about what dysfluencies a passage contains. In the following sections, the design of a procedure for the assessment of stuttered dysfluencies is described. The central feature of this is that it can be realized computationally in addition to being capable of being performed by human judges. The other main aspect of the design is its modularity that allows the system to be implemented as a whole or just selected parts. Later in this article, human judgmental data are to be collected according to the outlined procedure. The procedure results in improved inter-judge agreement about stuttered events and allows computational implementation of repetition and prolongation recognizers (described in Howell, Sackin, & Glenn, in press).

Considerations in designing the two-stage recognizer

The distinguishing characteristic of the procedure developed is that it involves processing the speech in two separate stages (c.f. Howell, 1978). In the first stage speech is segmented into linguistic units. Categorization of the type of dysfluency takes place in the second stage.

Segmentation

A distinction is drawn between dysfluencies that occur on single words (lexical dysfluencies, LD) and those that occur over groups of words (supralexical dysfluencies, SD). The types of dysfluency that are LD are prolongations (Ps), part-word and word repetitions (Rs), and broken words (Os). The remaining dysfluencies in Johnson et al.'s (1959) list are SD[1]. This distinction has ramifications for choice of the type of linguistic unit that needs to be located at the segmentation stage. Words were chosen here because they allow the division between LD and SD according to the earlier definition and, more importantly, stipulate the nature of the processing at the categorization phase for LD and SD types. Properties within words typify LDs,and SDs have characteristic patterns occurring over adjacent words (Howell, Au-Yeung, Sackin, & Glenn, 1997). Though word segments are used here, the word is not the only unit that is appropriate; in future work syllables will be used (see the discussion).

Categorization

The goal of categorization is an unequivocal assignment of each word as fluent, as a type of LD, or as part of an SD. The procedures that are currently used for clinical assessments do not meet this requirement as can be illustrated by considering Kully and Boberg's (1988) study. These authors sent copies of the same tape to different clinics for assessment. Consequently, they made no attempt to specify how judges ought to approach the assessments (this was, of course, prohibited by the research question they wanted to address). As a result, a section of speech that contains a phrase revision and a part-word repetition as in “..the boy c.came, the boy went …” is ambiguous; this dysfluent stretch could be counted as either or both of these types of dysfluency. When different judges categorize such a dysfluency in different ways, it would not only lead to poor interjudge agreement but the number of words considered “dysfluent” depends on the way the judges categorize the dysfluency too. This same problem will occur when automated procedures are developed for single dysfluency categories if no forethought is given about how to avoid

[1] Interjections are regarded as SD because they can be multi-word and because they often appear as components in other classes of SD.

such categorization ambiguities. The source of the confusions described is between dysfluencies in SD and LD classes. The strategy adopted both for human (current article) and automatic assessments (Howell et al., in press) to avoid this problem is to divide the categorization stage into two sequential phases. In the first phase words that are part of SD (whether they contain an LD or not) are located and processed to determine the SD fluency class. If a sequence of words has been designated into one of the SD categories, they are ignored in the second phase. Consequently, the words that remain are fluent or contain dysfluent words that are uniquely LD. This strategy gives precedence to SD over LD.

Computational implementation of components of the two-stage procedure: Strategy to by-pass components

One way of developing the dysfluency recognizers would have been to attempt lexical identification and then look for incomplete or repeated lexical patterns. Though this would involve the extra lexical identification step, it would have the advantage that pre-segmentation would not be needed. Though this is an advantage, note that the requirement of lexical identification is the factor that precludes this approach. At present, automatic recognition of even fluent lexical events is not robust enough so it can be employed on spontaneous speech, particularly that of children.

This led to the development of the strategy introduced in the previous sections that has the three essential components; Stage 1, segmentation of the speech; Stage 2, phase a, recognition of SD; Stage 2, phase b, recognition of LD. Lexical identification will not be required during any of these components so it will work on spontaneous speech as well as read speech. Thus, segmentation will be based on acoustic properties when the system is fully implemented. Acoustic patterns across words are mapped directly onto SD (phase a) and acoustic patterns within words are mapped directly onto LD (phase b).

A potential danger in designing a program of work intended to automatically identify all stuttered dysfluencies is that the system may need to be implemented as a whole before it can be assessed. However, by identifying the different modular components of the computational implementation this allows either human judgmental data to be used temporarily, or ad hoc computational solutions to be made, so that selected components can be skipped over. Both these strategies have been used when developing LD recognizers. Segmentations obtained from human judges (reported below) are employed in the development of the LD classifiers in Howell et al. (in press). Thus, the LD recognizer does not have to wait until a segmenter is perfected. Words that are part of SD are parsed out using a procedure that requires a transcription to be provided. The SD words that are excluded by the parser before human judgments about isolated LDs are made can also be excluded before judgments by the automatic recognizers are made. This solution is temporary insofar as, as discussed above, location of SD will ultimately not depend on the lexical units having been identified (some possible ways of developing automatic segmenters and SD recognizers are outlined in the discussion).

Information to be supplied from the psychometric procedures to meet the goals of LD categorization for R and P

The by-passing of Stage 1 and phase a of Stage 2, just described, will allow the next article to focus on developing automatic recognizers. The recognizers that will be developed are based on artificial neural networks (ANNs, Howell & Sackin, 1995; Howell, Sackin, Glenn, & Au-Yeung, 1997). The ANNs are trained during a supervised learning phase where information about instances of the dysfluency types that are to be recognized are presented. The ANN recognizers Howell et al. (in press) have developed are for word- and part-word-repetitions (R) and prolongations (P). Currently the R ANN does not distinguish between word and part-word repetitions, unlike in Johnson et al.'s (1959) psychometric assessments, but recognizes instances of both as R. The reason for this is that the detection of Rs at the start of words is similar whether the repetition is of the whole word or only part of it. This applies whether a human or a machine is doing the judging. The confusability between whether a repetition of an open syllable on an open-syllable word is a part-word, or word repetition could lead to disagreement between judges in a similar way to that occurs between SD and LD. This potential source of ambiguity is removed if these types of repetition are not differentiated. Though dealing with Rs as a class is the way that has been chosen to deal with this problem at present (another possibility would have been to remove them in SD parsing), some division of part- and whole-word repetitions will be necessary in the future. Thus, Throneburg and Yairi (1994) and Yairi and Lewis (1984) have reported that objective analysis of part-word repetitions, but not word repetitions, can help differentiate young children who are at the onset of stuttering from normally-speaking peers. The next step towards differentiating word- and part-word-repetitions will be to develop ANN recognizers for word repetitions on closed syllable words.

The ANN recognizers developed at present will not locate all categories of LD in Johnson et al.'s (1959) typology. In particular, no ANN recognizer for “broken words” has been developed. Wingate (1988) notes that this type of dysfluency is used to allocate words that do not fall within any other one. This, in turn, leads to the words that fall into this category being vaguely specified and heterogeneous in their properties. One consequence of this is that it is expected that interjudge agreement will be low for this dysfluency type. Development of ANNs for this category will have to wait until a closer specification of these dysfluencies (and, possibly, some subdivision of the category) has been made. In the psychometric assessments reported below, this class of dysfluency is designated “other” (O) dysfluency rather than “broken word” because broken word is a specific sub-type of this residual LD category. The broken word, proper, subtype does appear to have acoustic properties that may be used to differentiate it from R and P LD types, as well as from other subtypes of the O category. Thus, broken words per se appear to have hard attacks and follow a period of silence.[2]

A superficially similar pattern to that described on broken words, may occur during silent Ps . Alternatively these may be regarded as a sub-type of the P category. In silent Ps, there is a period of relative quiet followed by a hard attack when the speaker releases the sound. If the period during which the sound is prolonged were absolutely silent, then they would appear to share properties with broken words. However, in our experience, Ps are rarely, if ever, “silent”. When ANN recognizers are available for these subtypes, empirical investigation will settle this issue by establishing whether silent Ps are detected by P ANNs or by broken word ANNs.

Though the ANNs that are developed in Howell et al. (in press) are restricted to the gross R category and to Ps, they are nevertheless important. So, for instance, R and P provide information about whether stuttering is worsening. Stuttering is considered to be more acute when there is an acceleration across a sequence of Rs (van Riper, 1982) and when the proportion of Ps relative to Rs increases (Conture, 1990).

[2] The ANN strategy outlined in Howell et al., (in press) involves separate segmentation and classification stages. An important feature of the ANNs is that pauses are detected and available for inspection or for statistical analysis at the segmentation stage. As is apparent in the text, the occurrence of pauses is an important feature of several of the SD and LD category types. For this reason, silent pauses are not classified as a category of dysfluency in their own right but as a feature that is important for locating several types of dysfluency category.

Ancillary information about R and P that will be collected

Now that the formalization of the decision processes and justification for concentrating on the R and P categories are in place, the assessments that it is necessary for human judges to make can be described. Three basic steps have to be taken before the ANN software can be developed and tested. First, it is necessary to know how accurately boundaries between words can be located. Second, a procedure has to be included to remove words that are part of SD. Finally, the accuracy with which different judges can identify the fluency or LD dysfluency category of each word has to be obtained. The accuracy of word segmentation markers provided by different judges and that of a computational parser to remove SD words are described in the method. The word markers are used to select words to play to judges that make an LD dysfluency categorisation. As reasoned above, improved interjudge agreement over that reported by other investigators (Kully & Boberg, 1988) should result from the way the assessment procedures have been formalized.

More information about the origin of inter-observer disagreement is afforded by taking a rating about how fluently the word is judged to flow, as well as the categorisation of the word as F, P, R or O, LD types. Most assessments of fluency are based on the basis of LD category counts alone. However, it is usually considered that stuttered speech varies on a continuum from completely fluent to markedly dysfluent (Starkweather, 1987).

The flow ratings allow the judges to indicate where they would place a word they have categorized on such a continuum. The detailed information provided by the flow ratings should be particularly informative about Ps. Ps are likely to depend, mainly, on a durational judgment where the underlying dimension is continuous. For this reason, Ps are likely to be difficult to judge and variable between judges. This contrasts with other dysfluencies that may have properties anchored in a particular feature and, consequently, where it is comparatively easy to say whether a word has the property or not (i.e., is or is not a member of some dysfluency category). So, for example, one judge might place a mildly-prolonged word into the P category and indicate his or her ambivalence by indicating that flow was good whereas another judge might designate the word fluent (F) but indicate it was not flowing smoothly. In such cases, both judges would have indicated their misgivings about calling the word F, albeit in different ways. If this is so, then the flow ratings will have an important role in the selection of instances of Ps for training recognizers. In contrast to the situation described for Ps, Rs are likely to be detected based on the presence or absence of repeated sections of energy that is likely to be a categorical feature. Further consideration of the relevance as to how the agreement patterns across flow ratings impact on selection of training instances of ANNs is deferred until the data have been reported.

Method

Materials

The data used for speech assessments are from 12 children who stutter. The age of stuttering onset, history of any previous therapy and ages are summarized in Table 1. The children had been assessed by therapists and admitted on to a two-week intensive therapy course (described in Rustin, 1987).

Recordings were made at the commencement of the course (i.e., before they had received this particular therapy program). Several types of speech sample were obtained but attention is restricted here to a reading of the 376-word passage "Arthur the rat" (Abercrombie, 1964, adapted from Sweet, 1895). Recordings of the children who stutter were made in a sound-attenuated booth. The text was pinned to the wall of the booth in line with the eyes. It was positioned at a comfortable distance for reading. The speech was transduced with a Sennheiser K6 microphone positioned six inches in front of the speaker in direct line with the mouth. The speech was recorded on DAT tape and transferred digitally to computer for further processing. The speech from the DAT tapes was down-sampled to 20 kHz.

Judges

Transcriptions, word-boundary and fluency-category judgments were made independently by two (male) judges. Both judges have considerable experience in marking word boundaries and in categorizing stuttered events (six years and two years, respectively) but not with the procedure described here.

Parsing applied to remove SD

Transcriptions were made by the two judges independently. These were in orthographic form with word attempts not indicated e.g., k..k..Katy would be Katy. Between the two transcribers, 95% agreement was achieved for all speech material used. The transcription of a speaker's reading was compared against a coded form of the text. To produce this coded form, the ordinal position of each word in the text was noted. A dictionary of the words and their position tag was created (referred to as the master list). In this dictionary, a single word entry can have more than one word position (e.g., common words like “and”).

The transcription of the passage was first processed by retrieving the position or positions of each word in sequence from the master list. Any word that did not occur in the master list was tagged with an arbitrary high value, 9999. The program next checked and eliminated inappropriate word positions for the word entries that had multiple positions. Word positions were eliminated using the position of neighboring words as a constraint. The sequence of words “ .. dead half in and half out ..” (used in all subsequent examples) has the word positions indicated in brackets: dead (351), half (352, 355) in (102, 353) and (45, 114, …, 354) half (352, 355) out (31, 155, 356) This would become: dead (351) half (352) in (353) and (354) half (355) out (356)

A word list with continuous word positions, as above, indicated that no SD was present. Conversely, a discontinuous word position signaled that words had been inserted or missed and, therefore, that an SD had occurred. An inspection, by the program, over the near continuous word positions then determined the nature of the SD.

For revisions embedded within a repeated phrase (a subset of category b), the original word sequence and associated positions might be: dead (351) half (352, 355) in (102, 353) sorry (9999) half (352, 355) in (102, 353) and (45, 114, …, 354) half (352, 355) out (31, 155, 356) After elimination of inappropriate word positions, this becomes: dead (351*) half (352*) in (353*) sorry (9999) half (352) in (353) and (354*) half (355*) out (356*)

The word positions with a “*” are located as the nearest continuation of the word position sequence. Thus, a jump back over several previous continuous words with an extra word or words (after the word “sorry” in this example) indicated a revision (category b). A phrase repetition (category d) occurred when this pattern occurred without an extra word or words. An example of an interjection (category a) would be: dead (351) or (9999) half (352, 355) in (102, 353) and (45, 114, …, 354) half (352, 355) out (31, 155, 356) After processing for adjacent words, this becomes: dead (351*) or (9999) half (352*) in (353*) and (354*) half (355*) out (356*)

Thus, a jump over several words but with the words otherwise in a continuous sequence with an extra word or words that was not found in the immediately preceding context may have indicated a word insertion (category a). This same pattern can also occur in revisions out of phrasal context (another subtype of category b where there are no associated word repeats). An example of an incomplete phrase (category c) would be: dead (351) half (352,

355) and (45, 114, …, 354) half (352, 355) out (31, 155, 356) This becomes: dead (351*) half (352*) and (354*) half (355*) out (356*) So a forward jump over one or more words but with the words otherwise in continuous order indicated a word or several words was/were missed.

The words that were part of SD were repaired by the parser so that they were removed before statistical analysis of LD was made. This was done by marking words corresponding to the first sequence of duplicated continuous numbers and any inserted words that were located. One or more repetitions of a complete word were marked so that they remain during statistical analysis because they are considered as instances of LD. This strategy gives precedence to SD over LD. The removal of inserted words and phrases that are repeated subsequently in SD (in the example given in the introduction, for instance, the words “the boy c.came”), prevents SD-LD confusions like those discussed. This is because the speech that remains only contains LD that are not part of SD. Also, the speech that is left is close to the original text (the discrepancies that remain are due to omitted words still being absent).

Word Boundary Judgments

The judges had to locate all the words noted in the transcription of the passage of each child who stuttered. The judges had to locate the point at which each word ended. This ensured that any pauses or word-initiation attempts were included at the start of the subsequent word (for the purpose of the current experiment, word repetitions were also treated as word-initiation attempts). The position where each word ended was located with the aid of two travelling cursors that were superimposed on the oscillograhic display of the speech waveform. Each cursor position was independently adjusted with a mouse. The first cursor was initially placed at the start of the speech. The second cursor was positioned slightly before where the end of the first word was thought to be located. The section of speech between the first and second cursors was then played and listened to over an RS 250-924 headset, to check whether the end of the word had been correctly located. The second cursor was then adjusted further forward and the word replayed again. Once the judge had adjusted past the end of the word, he adjusted the second cursor back in time into the file until the end of the word was not heard. By continually crossing backwards and forwards across the boundary at the end of the word, he became satisfied that he had located it accurately. This marker was stored and the subject proceeded to locate the boundary between the next words in the sequence in the same way (this procedure is similar to that used by Osberger & Levitt, 1979). It was found necessary to include part of the silent gap in fluent words that ended with a plosive (e.g., the t of “cat” in “catflap”) as the silent interval is a salient cue for perception of the plosive. Filled pauses were treated as words both here and when the words were categorized. The judge worked through the recordings of each speaker sequentially.

Assessment Procedures for Dysfluencies

After the word-endpoint markers had been obtained, each judge made his fluency assessments. Words were defined as the interval from the end of one word to the end of the next. The analyses reported below show that the word-endpoint marker judgments are not significantly different. Consequently, only one set need be used. The markers of the more experienced judge were employed to select the words for assessment. Each word in the speech of all 12 participants (including the words located by the parser as part of an SD), was judged once by each judge.

The test procedure required specifying a random presentation order for all words. A randomized presentation order was used so that the global context in which all judgments were made was as constant as possible (thus, minimizing range effects, Parducci, 1965). The first randomly-selected word was heard in isolation. After a short pause the test word was heard along with the word that had preceded it (the two words had the same timing and were in the same order as in the original recording). Consequently, pauses were apparent when they occurred between the context and test word. The test word alone and this word with the word that preceded it could be heard as often as the judge required by hitting the return key on the computer keyboard. Thus, presentation of the test and context words was initiated by the judge, and the current trial ended and the next trial commenced after the judge entered his responses (detailed in the following paragraph). The full context was not available at the beginning of the recording (in these cases, the word to be judged was still played in isolation beforehand). It was stressed to the judges that judgments were being made about the first word that was played.

The two judgments obtained about each word, were a rating about how comfortably the word “flowed” and a categorisation of the word as F, P, R or O. The flow judgments were made after the categorisation on a 5-point scale. The 5-point scale employs a Likert scale format (Likert, 1932) in which the judge indicated the extent to which he agreed with the statement that “the speech is flowing smoothly” (1 = agree, 5 = disagree with intermediate values showing intermediate levels of agreement). Thus, it represents a judge's assessment of the speaker's ability or inability to proceed with speech (c.f. Perkins, 1990). It was stressed that the rating scale was not a finer-grained indication of whether a word is fluent or dysfluent: A word high in “flow” might nevertheless be categorized as F or vice versa. Judges were self-paced in that they ended the judgment session when they began to feel fatigued. On recommencing, they started at the point in the random sequence where they had left off. Typically judges would judge about 200 words before taking a rest.

Results

Word Boundary Location

The average (unsigned) difference in boundary placement between judges was 14 ms (SD = 8.6 ms) for fluent words (3.2% of the average of these words' duration), and 42 ms (SD = 24.1 ms) for stuttered words (4.7% of the average of these “words'” duration). Judge 1 placed his marker before Judge 2's on 48.7% of the words, Judge 2 placed his marker before Judge 1's on 44.2% of the words and the two judges placed their markers at exactly the same spot on 7.1% of the words (all based on segmentation markers to the nearest ms). Because Judge 1 was about as likely to place his marker in advance of Judge 2 as the reverse, it appears that the (small) differences that occurred were due to chance.

Lexical Dysfluency Categorisations

Words identified as parts of SD that should be removed to repair the speech were excluded during analysis. Each remaining word was classified with respect to the category assigned to it by Judge 1 and the category assigned it by Judge 2 (F, P, R or O for each judge). The counts over all words for F, P, R and O categories are presented in Table 2. The convention used to refer to individual cell entries is to use the first letter to stand for the categorisation given by Judge1 and the second letter stands for that given by Judge 2. Most words occur in the FF category.

Agreement was calculated for each dysfluency category by taking the sum of agreements about the selected category divided by the sum of these agreements plus all disagreements and converting them to a percentage. This is a strict criterion as it excludes the FF category from agreements whilst at the same time allowing responses that one of the judges placed in the designated category but the other judge made another response to (including F) to count as a disagreement. Agreement was 24.3% for Ps (82/337), 76.7% for Rs (181/236), and 19.1% for Os (56/293). Agreement calculated for F words was 89.9% (3700/4115) and for words overall 89.3% (4019/4500). Disagreement about type of dysfluency D, calculated according to the same criterion was relatively low at 17.1% (66/385). Thus, agreement was satisfactory for the overall data, for Rs, and for Fs. Agreement was unsatisfactory for P and O words. The poor agreement for Ps is of note because the assessments were intended to provide agreed instances of P for training ANNs. The poor agreement for Os is less crucial as it is not intended to develop ANN recognizers for this class of dysfluency at present. Inspection of the top row and first column of the matrix in Table 2 shows that Judge 2 was more inclined to use a D category for a word than Judge 1 (a large number of words that Judge 1 called F, row one, were placed into one or other D category by Judge 2). Though Judge 2 designated more words D, he may have, as hypothesized in the introduction, reflected his uncertainty about giving a word a D categorisation using the flow ratings. To assess this, the data were broken down by his flow ratings and these data are shown in the first five sections of Table 3.

It can be seen in the matrices that for the group of words given a rating of 2, shown in Table 3, Judge 2 categorized 82 words as P that Judge 1 categorized as F. At this rating, the corresponding effects for the two other dysfluent categories (R and O) were far less dramatic. Similarly, disagreement about Ps for words given flow ratings of 3 was high (this now also applied to Os for this group of words, though still not to Rs). For Ps with higher flow ratings, there was an abrupt decrease in disagreements for Ps but not for Os. Considering Ps first, it appears that Judge 2 adopted the strategy of calling some marginal words P and registered his uncertainty by giving those words a low flow number. For the group of words given a two flow-rating, Judge 1 categorized no words as P. Judge 1 started to call three-flow rated words P, but (a) there were far fewer of these than for Judge 2, (b) a small proportion, relative to the other judge, was called P when the other Judge called them F. A Chi square test on the F × P cells across judges for the three-rated words indicated that the marginal totals were not significantly different from chance distribution (Chi square = 0.02, df = 1, p<0.05). In other words, the two judges did not seem to be affected in the same way by the characteristics of the stimuli for the group of words given 3 flow rating by the experienced judge but, rather, appeared to be guessing.

These difference between the judges in making P judgments may be due to an inherent subjectivity in the way duration was used to make judgments about Ps. Judges have no fixed yardstick for deciding how prolonged a sound must be to be called “P” and, consequently, each judge might legitimately have adopted a different duration criterion to establish whether a sound was prolonged. If this was so, then there should be systematic differences in duration for the words in the F × P cells.

The mean duration of the words in the FF cell (0.306s) was less than that in the PP cell (0.968s). This supports the contention that duration is an important feature, though not the only one as Howell et al. (in press) show, for differentiating agreed F words from agreed P words. Any systematic differences between judges in the duration criterion used, would be revealed by asymmetries in the average duration of the PF and FP cells where only one of the judges considered the sounds were prolonged in these cells. The average duration of the words in the FP cell was 0.535 s whereas that in the PF cell was 0.672 s. The Z score for the difference between the two means was 2.5 (p<0.01). Thus Judge 2 was calling words of shorter duration prolonged than was Judge 1. This finding is consistent with the fact that Judge 2 found more Ps, as he would have included more words with shorter durations in this dysfluency category. The P words given flow ratings of 4 and 5 had sustained phonemes that led both judges to designate them P (the average duration of the 40 agreed P words at a flow rating of 4 was 0.727 s and that of the 28 agreed P words at a flow rating of 5 was 1.505 s).

The P words given a 4 or 5 flow rating would be candidates for selection as agreed instances provided that interjudge agreement was acceptable. For words given 4 and 5 ratings alone, agreement was 80% for Ps (68/85), 85% for Rs (121/143), and 43% for Os (24/56) calculated in the same (strict) manner as before. Overall agreement was 82% (213/259) and disagreement about dysfluencies 11% (25/238). Agreement about each separate dysfluency category improved, particularly for Ps. As expected, agreement for Os was still not high enough so that reliable training instances could be selected even for 4- and 5-rated words alone. A further feature to note is that these agreements exclude any biassing due to the high number of agreed F words (the FF cell had no entries for words given 4 and 5 ratings).

Discussion

In the introduction, three components of a strategic plan to automatically locate stutterings were outlined. These three components, in brief, are segmentation of the speech, location and removal of SD and location of the LD that remain. This article has primarily concentrated on location of reliable instances of LD by human judges but, to get into a position to address this, the earlier components have had to be considered too. Word markers were annotated on the speech. Positioning of the word markers by two judges was close, irrespective of whether the material was F or D. Furthermore, there was no systematic biases for one judge to position his markers in advance of the other. This allowed selection of either set of markers for subsequent psychometrical assessment. As argued, words that are part of SD have to be removed prior to LD-location as, otherwise, they can lead to ambiguous category affiliation of a word. As a temporary solution to dealing with SD, words that were part of SD were located and removed by a parser. This used a transcription of each reading as input. This is a temporary solution because a feature of the strategy that has been planned is that lexical identification should not be required during any of the components. Contrary to this requirement, the transcription provides an indication of the lexical identity of the words in the passage as read. After words that were part of SD had been marked so that they can be ignored during statistical analysis, words were assessed for selected LDs. In the LD assessments, the words selected for test were based on the markers of the most experienced judge.

Elsewhere, agreement between judges about stuttered events has been reported to be poor (Kully & Boberg, 1988). A logical consequence of formalizing the strategy so that SD-LD ambiguities do not occur should be an improvement in inter-judge agreement. Overall agreement was 89.3%. This is likely to be near the upper limit of what is achievable, in the light of the evidence for guessing on Ps of words rated as 3 by the experienced judge. Direct comparison with other literature is made difficult given the differences between the current procedure and others that are adopted, reports of statistical quantities (rather than % agreement) and the predominant use of overall agreement rather than event-by-event measures (the latter are equivalent to the type of agreement reported here). An early study by Johnson and Colley (1945) does provide some basis for comparison. In this study, two judges identified stuttering in connected speech by depressing a key at the beginning of each moment of stuttering and releasing it at the end of the moment. The moments of stuttering would include SD as well as LD. Event-by event agreement between the two judges was 72.3% (1427/2315), some 17% lower than obtained here.

As stated earlier, the main reason for assessing agreement about categories across the two judges was to provide reliable instances for training and testing ANNs designed to recognize Rs and Ps. It was necessary, therefore, to assess agreement for these specific LD (rather than overall agreement concerning F or D, as discussed earlier). The results for unselected Rs showed reasonable interjudge agreement at 76.7%. Agreement, at 80%, for Ps was only satisfactory for words rated 4 or 5 by the experienced judge (the agreement on the equivalent R material was 85%).

There are several potential ways these results could be used to select reliable instances of Ps and Rs for training ANNs during supervised learning: The two most obvious of these are, (a) to pick materials that both judges agree about the dysfluency category, or (b) to use criteria that employ rating and categorization responses obtained from a single judge. In the following paper, the second option was chosen. However, a criterion is applied to the rating and categorisation responses in a way that goes some way towards obtaining material about which both judges agreed (in part fulfilling the intention behind the first option). The criterion is to select the words categorized as dysfluent and which were given a 4 or 5 flow rating by the most experienced judge (inter-judge agreement was 80% or above for Ps and Rs). An advantage of selecting the responses of the most experienced judge is that this was also the judge whose word boundaries were used for presenting words in the LD test. Nevertheless, the selection criterion is a compromise insofar as it allows instances to be selected that show good interjudge agreement while at the same time going some way down the road of modelling the responses of a single judge (which is the preferred approach for the reasons given in the next paragraph). It implicitly acknowledges that poorer agreement occurs between judges on the instances that have been omitted.

In subsequent phases of this project, it is intended to investigate further these and other ways of pooling responses to obtain reliable instances of dysfluency categories across different judges. The intention in pooling responses is to obtain some consensus about whether particular instances are from one dysfluency category or another. The most straightforward way of doing this is to have a panel of judges and to ascertain what dysfluency class receives the majority verdict on each word. These could then be used to select representative instances and a single ANN representing a composite judge could be trained. One caution that needs to be raised is that this is only one way of selecting training instances for a composite ANN. The input to the composite ANN might be words selected so that they are designated in a D class and given a 4 or 5 flow rating (as here) or there might be no preselection at all. Each of these alternatives may be deemed important depending on the research application: Selection of the training instances using ratings was considered important here in order to achieve a measure of comparability across judges. As has been pointed out, this is equivalent in some respects to selection of instances that a number of judges give the same dysfluent category response to the word. Such selection might be inadvisable when it is important not to miss any stutterings even at the expense of allowing some false alarms. A practical illustration of this would be if it was important not to misdiagnose a speaker who stutters as a fluent speaker.

The reason that training of ANNs for individual judges, rather than a composite judge, has been chosen is as follows. The ANN outputs modelled on the responses of each individual judge are not discrete responses as they are with human judges (representing in the above experiment a vote for the word falling into a particular dysfluency class). Consequently, the outputs of each individual-judge ANN, potentially contain additional information that can be employed to improve pooled decisions over the decisions obtained when discrete class judgments from human judges are pooled and used as input to a single composite ANN. To illustrate, ANNs can be trained for individual judges that have outputs (as in Howell et al., in press) representing R or P. The ANNs produce an output on a continuous scale that represents the likelihood of each word being R or P or, if there is no evidence that either of these are likely, the word is F. The separate outputs of ANNs representing a panel of five judges might show that a word is called P by two ANN-judges and F by the remainder. The discrete responses of the ANNs would appear to favor calling the sound F. However, the picture might change depending on the level of activation for P in the three ANNs that designated the sound F. Rather than basing the decision on an arbitrary rule about how to pool the outputs of individual-judge ANNs, the pooling could be based on a second layer neural network.

Assume that individual ANNs are trained for each member of a panel of judges and each word has a categorisation and activations for each output. These provide a set of input parameters that could be employed to train the second level ANNs that pool responses in different ways. For these second level ANNs, some method of establishing what output is appropriate has to be specified. This could be a criterion of a majority of the judges giving the word the same dysfluency categorisation. Alternatively, it could be any judge giving the word a dysfluency categorisation, one selected judge, and so on. Once the output has been unequivocally specified in one of these manners, then the second layer network can take all the responses and all the activations of each word by the ANNs trained for each member of the panel of judges. The second layer ANN will then learn what is the best balance of information provided by the input parameters to arrive at the specified response. Thus for instance, even when a criterion of “majority of judges” is employed, the network might learn to give a low weighting to the responses of a judge who has a low criterion for assigning words to a dysfluency category. Such a judge would produce a lot of dysfluent responses in addition to those that are associated with a majority verdict. The additional advantage preferred by this approach is that it permits flexibility at the second level as just described. So, for instance, the bank of ANNs can easily be extended when a new judge has to be added. To do this, ANNs would need to be trained for that judge and the second network retrained to incorporate the judge. However, the extant ANNs for other judges would not need retraining. The procedure outlined for training judges is easily extendable to judgments about single speakers as opposed to a group of speakers.

The other feature shown by these data that has a bearing on the design of the ANNs is, as occurs elsewhere, the gross imbalance between occurrences of F words and types of words assigned to a D category. This poses potential problems for training ANNs: The ANNs could quickly learn to categorize everything F that would give a high overall level correct but show no sensitivity to Ds. Consequently, the training material has to be chosen selectively and some attempt needs to be made to balance the composition of the training data with respect to instances of F and D categories.

The study has provided the information that is necessary for the LD classification stage of the ANN recognizer (Howell et al., in press). The properties of the segmentation and SD categorization phase were pragmatically-motivated to this end. A question that arises about both these components is how feasible is it to implement them computationally (a prerequisite to a fully automated version)? As stated in the introduction, the crucial thing about automating segmentations is that they need to be based on acoustic characteristics. Exploratory investigations have been made on use of a dynamic time warping algorithm (Deller, Proakis, & Hansen, 1993) with promising results (Howell, Sackin, et al., 1997). The main drawback in using dynamic time warping to obtain word boundaries is that it requires a dictionary that contains templates of the words that are to be segmented. Therefore, it can only be employed with read texts and, ideally, a fully automated version will work on unrestricted (i.e., spontaneous) speech.

One speech unit that it is possible to locate automatically in unrestricted speech is the syllable. Syllables have associated acoustic properties in that they contain a vowel and, consequently, have a single intensity peak associated with them.[3] An important reason for

[3] The text employed in this experiment consists mostly of monosyllables so the current results offer an idea of what performance will be like with syllables too.

choosing to use syllables is that their acoustic properties are more robust in the vicinity of stutterings than they are in fluent stretches of speech (Howell & Wingfield, 1990). Consequently, the syllables that it is crucial to locate to establish whether they are D are likely to be easiest to segment accurately. Another factor that makes it comparatively easy to locate stuttered syllable segments is that these are likely to be stressed (Wingate, 1988) that makes them acoustically more prominent. Indeed, it may be desirable to establish whether each syllable is stressed or not as this provides important input parameters that are necessary to recognize SD (Howell & Young, 1991). Because syllables will be located directly from their acoustic properties, the identity of the syllables does not have to be established. Therefore, it is possible to use these segments on spontaneous speech. Syllables have the drawback that they split Rs up so that their properties occur over adjacent syllable segments. However, this can easily be dealt with by looking for patterns across the syllables in a similar way to that described for SD in the next paragraph. Syllables also have properties that commend their clinical use. Thus, most current literature reports measures of stuttering as percentage of syllables stuttered or number of stutterings per 100 syllables.

Turning now to what will be necessary to acheve automatic SD location, first consider what the parser does. The parser knows the target and produced word sequences and looks for discrepant patterns over what are expected. In the complete implementation, the identity of the target and produced words will not be known (nor, obviously, their constituent syllables). A fully automatic version could locate similar SD patterns from acoustic inputs that the parser detects in lexical strings. If the words “the boy c.came, the boy went” were spoken, a technique like dynamic time warping could be used to establish the acoustic similarity between the words in the sequence. This would reveal that the two “the”'s and the two “boy”'s are acoustically very similar and “c.came” is different from “went”. If the parser was supplied with this information, it could locate the words in an equivalent manner to the way it currently does (based on a known sequence of words). Incomplete phrases could not be located in this fashion. Ancillary information about this class of D as well as of other SD could be provided from pauses, syllable stress and, potentially, other prosodic information such as speech rate and intensity. Though automatic location of SD is likely to be difficult to achieve, it deserves attention because these types of dysfluency are receiving renewed attention (Howell, Kadi-Hanifi, & Young, 1990; Lasalle & Conture, 1995; Postma & Kolk, 1993).

In summary, a strategy has been outlined for automatically locating stuttered dysfluencies. This involves the two stages of segmentation and classification (with classification taking place in two phases). Care has been taken to ensure that all compnents can potentially work on acoustic input alone (so as to allow development of a fully automatic version that is applicable to spontaneous speech). The experiment with human judges, reported in this paper, involves all these components though, admittedly, the SD classification phase involves knowledge of word sequence and cannot be said to be based on acoustic inputs. In one sense, the experiment simulates the entire process with human judges. However, it has more significance than pure simulation. The modular design allows one or more components to be left out during this development phase. The results on word segmentation and location and exclusion of SD allow the next paper in this series to address LD classification directly. The procedures for selecting R and P instances that judges agree about can be employed to provide suitable materials for obtaining input parameters for ANNs under a supervised learning regime. The information on overall R and P categorization can be used for testing the ANNs and comparing performance against human judgmental data.

Acknowledgments

This research was supported by a grant from the Wellcome Trust. Grateful thanks to the children, their parents and the staff of the Michael Palin Centre for Stammering children (particularly Lena Rustin and Frances Cooke). The comments of Drs. Au-Yeung, Curlee, Colcord and Jennifer B. Watson are gratefully acknowledged.

References

Abercrombie, D. English phonetic texts. London: Faber and Faber; 1964.

Conture, EG. Stuttering. 2nd Edition. Englewood Cliffs, New Jersey: Prentice-Hall; 1990.

Deller, JR.; Proakis, JG.; Hansen, JHL. Discrete-time processing of speech signals. Englewood Cliffs, New Jersey: Macmillan; 1993.

Howell P. Syllabic and phonemic representations for shortterm memory of speech stimuli. Perception and Psychophysics. 1978; 24:496–500. [PubMed: 750990]

Howell P, Au-Yeung J, Sackin S, Glenn K. Detection of supralexical dysfluencies in a text read by child stutterers. 1997 Manuscript submitted for publication.

Howell, P.; Kadi-Hanifi, K.; Young, K. Phrase repetitions in fluent and stuttering children. In: Peters, HFM.; Hulstijn, W.; Starkweather, CW., editors. Speech motor control and stuttering. New York: Elsevier; 1990. p. 415-422.

Howell, P.; Sackin, S. Automatic recognition of repetitions and prolongations in stuttered speech. In: Starkweather, CW.; Peters, HFM., editors. Proceedings of the First World Congress on Fluency Disorders; Nijmegen, The Netherlands: University Press Nijmegen; 1995. p. 372-374.

Howell P, Sackin S, Glenn K. Development of A Two-Stage Procedure for the Automatic Recognition of Dysfluencies in the Speech of Children Who Stutter: II. ANN Classification of Repetitions and Prolongations with Supplied Word Segment Markers. Journal of Speech and Hearing Research. in press.

Howell, P.; Sackin, S.; Glenn, K.; Au-Yeung, J. Automatic stuttering frequency counts. In: Peters, HFM.; Hulstijn, W.; von Lieshout, P., editors. Speech motor control and stuttering. New York: Elsevier; 1997.

Howell P, Wingfield T. Perceptual and acoustic evidence for reduced fluency in the vicinity of stuttering episodes. Language and Speech. 1990; 33:31–46. [PubMed: 2283919]

Howell P, Young K. The use of prosody in highlighting alteration in repairs from unrestricted speech. Quarterly Journal of Experimental Psychology. 1991; 43:733–758. [PubMed: 1775664]

Johnson, W.; Boehlmer, RM.; Dahlstrom, WG.; Darley, FL.; Goodstein, LD.; Kools, JA.; Neeley, JN.; Prather, WF.; Sherman, D.; Thurman, CG.; Trotter, WD.; Williams, D.; Young, M. The onset of stuttering. Minneapolis: University of Minnesota Press; 1959.

Johnson W, Colley WH. The relationship between frequnecy and duration of moments of stuttering. Journal of Speech Disorders. 1945; 10:35–38.

Kully D, Boberg E. An investigation of inter-clinic agreement in the identification of fluent and stuttered syllables. Journal of Fluency Disorders. 1968; 13:309–318.

LaSalle L, Conture E. Disfluency clusters of children who stutter: Relation of stutterings to self-repairs. Journal of Speech and Hearing Research. 1995; 38:965–977. [PubMed: 8558887]

Likert R. A technique for the measurement of attitudes. Archives of Psychology. 1932; (140)

Osberger MJ, Levitt H. The effect of timing errors on the intelligibility of deaf children's speech. Journal of the Acoustical Society of America. 1979; 66:1316–1324. [PubMed: 500969]

Parducci A. Category judgment: A range-frequency model. Psychological Review. 1965; 17:9–16.

Perkins WH. What is stuttering? Journal of Speech and Hearing Research. 1990; 55:370–382.

Postma A, Kolk H. The covert repair hypothesis: Prearticulatory repair processes in normal and stuttered disfluencies. Journal of Speech and Hearing Research. 1993; 36:472–487. [PubMed: 8331905]

Rustin, L. Assessment and therapy programme for dysfluent children. Windsor, England: NFER Nelson; 1987.

Starkweather, CW. Fluency and stuttering. Englewood Cliffs, NJ: Prentice-Hall; 1987.

Starkweather, CW.; Gottwald, SR.; Halfond, MM. Stuttering prevention: A clinical method. Englewood Cliffs, NJ: Prentice Hall; 1990.

Sweet, H. A primer of spoken English. Oxford: Clarendon Press; 1895.

Throneburg RN, Yairi E. Temporal dynamics of repetitions during the early stage of childhood stuttering: An acoustic study. Journal of Speech and Hearing Research. 1994; 37:1067–1075. [PubMed: 7823553]

van Riper, C. The nature of stuttering. Englewood Cliffs, New Jersey: Prentice-Hall; 1982.

Wingate, ME. The structure of stuttering: A psycholinguistic study. New York: Springer-Verlag; 1988.

Yairi E, Lewis B. Disfluencies at the onset of stuttering. Journal of Speech and Hearing Research. 1984; 27:154–159. [PubMed: 6717001]

Table 1

Details of the speakers used

The age of onset, history of any previous therapy and age at recording are presented in columns three, four and five, respectively.

Subject	Age at onset	Previous therapy	Age at recording
RPi	3:6	Yes	11
RPe	2:6	No	10
MD	4:0	No	12
JP	8:0	Yes	12
BA	2:0	No	10
BL	4:0	No	10
MT	3:6	Yes	10
DN	4:0	No	12
WR	3:0	No	10
AR	4:0	No	10
JB	6:0	No	13
MC	2:0	Yes	11

Table 2

Confusion matrices Judge 1 (side) against Judge 2 (top)

In the top section counts of words assigned to F, P, R and O categories are shown. This matrix is decomposed into the separate F × P, F × R and F × O confusion matrices in the following section.

Overall					
			Judge 2		
		F	P	R	O
	F	3700	205	14	116
	P	8	82	2	3
Judge 1					
	R	2	13	181	20
	O	70	24	4	56
Overall agreement = 89.3%					

Table 3

Flow ratings

F × P × R × O matrices and separate F × P, F × R and F × R matrices represented in the same manner as in Table 2 for the separate flow ratings 1-5 and for flow ratings 4 and 5 combined.

Flow ratings 1					
			Judge 2		
		F	P	R	O
	F	1837	0	0	0
	P	0	0	0	0
Judge 1					
	R	1	0	0	0
	O	1	0	0	0
Flow ratings 2					
			Judge 2		
		F	P	R	O
	F	1587	82	1	6
	P	0	0	0	0
Judge 1					
	R	1	0	0	0
	O	25	5	0	0
Flow ratings 3					
			Judge 2		
		F	P	R	O
	F	276	121	9	96
	P	8	14	0	1
Judge 1					
	R	0	7	60	12
	O	43	14	2	32
Flow ratings 4					
			Judge 2		
		F	P	R	O
	F	0	0	4	14

	P	0	40	1	1
Judge 1					
	R	0	4	80	3
	O	1	3	2	22
Flow ratings 5					
			Judge 2		
		F	P	R	O
	F	0	2	0	0
	P	0	28	1	1
Judge 1					
	R	0	2	41	5
	O	0	2	0	2
Flow ratings 4 and 5					
			Judge 2		
		F	P	R	O
	F	0	2	4	14
	P	0	68	2	2
Judge 1					
	R	0	6	121	8
	O	1	5	2	24