Published in final edited form as:

J Acoust Soc Am. 1998 December ; 104(6): 3558–3567.

Methods of Interval Selection, Presence of Noise and their effects on Detectability of Repetitions and Prolongations

Peter Howell, Alison Staveley, Stevie Sackin, and Lena Rustin

Department of Psychology, University College London, Gower Street, London WC1E 6BT, England

Abstract

Accurate methods for locating specific types of stuttering events are necessary for diagnosis, treatment and prognosis. A factor that could add variability to assessment of stuttering is noise on recordings. The effects of noise were assessed by adding noise to intervals of speech containing all fluent material, fluent material with a repetition or fluent material with a prolongation. These intervals allow a unique dysfluency response to be made. A statistical analysis of the occurrence of such intervals in spontaneous speech showed only a limited number of intervals met these criteria. This demonstrated that selecting intervals at random from spontaneous speech (as in Time Interval analysis procedure) will infrequently lead to a unique and unambiguous dysfluency specification for the interval. Intervals were selected for test from the intervals that met the stipulated criteria. These were presented for dysfluency judgement when the position of the stuttering within an interval was varied and with different amounts of added noise (no added noise, 3 dB and 6 dB of noise relative to mean speech amplitude). Accuracy in detecting stuttering type depended on noise level and the stuttering's position in the interval, both of which also depended on type of stuttering: Noise level affected detection of repetitions more than prolongations: Repetitions were more difficult to detect when they occurred at the end of an interval whereas prolongations were more difficult to detect when they were at the beginning of an interval. The findings underline the importance of adopting rigorous recording standards when speech is to be employed to make stuttering assessments.

INTRODUCTION

Measurements of the incidence of stuttering in samples of speech are required for a number of purposes: They are used to aid clinicians decide who to treat (diagnosis), to assess what changes in speech occur after treatment (treatment outcome) and to help establish which individuals are likely to be treated most successfully (prognosis). However, the measurement procedures that are traditionally used in clinics produce variable estimates of stuttering when different judges assess the same samples of speech (Kully & Boberg, 1988). Some approaches being investigated to improve assessments are motor metrics (Alfonso, 1990; Smith, 1997), automatic procedures (Howell Sackin & Glenn, 1997a; Howell, Sackin, Glenn & Au-Yeung, 1997) and improved psychometric methods (Howell, Sackin & Glenn, 1997b; Ingham, Cordes & Gow, 1993).

The audio recordings of speech used for assessment should also be made to high standards. However, it is not always easy to achieve good quality recordings in clinical environments. The effects noise introduced during recording has on assessments have not been investigated previously. The following study was conducted to ascertain how such noise affects the detectability of repetitions and prolongations. These types of stuttered dysfluency were

selected as an increase in the proportion of prolongations to repetitions is an important diagnostic (Conture, 1990; Howell, 1993) and prognostic indicator (Conture, 1990). They can be reliably detected on words when recordings are made in controlled acoustic environments (Howell et al., 1997b). For reasons given in the next section, assessments of the effects of noise have to be made in fixed-length intervals. In the final section, background evidence concerning how noise affects detection of repetitions and prolongations is given.

A. Test interval selection criteria

Speech is listened to in continuous context when samples are assessed in clinics. Such procedures offer no control over whether judges give equal attention to all sections and whether speech in the surrounding context affects judgements about the stretch (Parducci, 1965). In the current study, experimental material was selected so: a. The response to the test extract is unambiguous (fluent throughout, contains a repetition, contains a prolongation). b. The material can be presented in equivalent contexts so that judgements are not biased. Though on the face of it these seem simple requirements, matters are somewhat complicated.

The first requirement is met by presenting speech for judgement that contains all fluent material, a single repetition or a single prolongation. Care has to be taken when selecting fluent material since some types of stuttered dysfluencies extend over groups of words (referred to by Howell, Au-Yeung, Sackin, Glenn & Rustin, 1997 as supralexical dysfluencies). Supralexical dysfluencies include phrase repetitions, phrase revisions and idea abandonments (Wingate, 1988). Words within supralexical dysfluencies can appear perceptually fluent on superficial listening and would be judged accordingly when presented in isolation. However, they have different prosodic structure relative to fluent speech that is not part of a supralexical dysfluency (Howell & Young, 1991). Supralexical dysfluencies were identified and marked so that the words that occur in them can be excluded when test material is selected. After this exclusion, the speech that remains has dysfluencies that occur within the bounds from the start of one word to the next (including repetitions and prolongations) usually on the first part of the first syllable in the word (Brown, 1945; Wingate, 1988). Dysfluencies like this are called lexical dysfluencies.

Syllables selected according to dysfluency type are presented in a fluent context to meet the second requirement, mentioned at the outset of this section. The reason a fluent context is necessary can be appreciated by considering what would happen if the syllable that contains a lexical dysfluency or a fluent syllable alone was presented. The syllables that contain a lexical dysfluency are usually longer than the fluent ones. The syllables are relatively easy to locate when noise on recordings is low. The duration of the syllables would be apparent when they are presented in isolation. The duration would remain apparent when noise is added to the syllable. Listeners could then use duration to decide what type the syllable is irrespective of whether noise on recordings prevents judges determining where a syllable starts and finishes in continuous context. Consequently, though it could be more difficult to extract syllable duration as information about dysfluency in noisy recordings, this would not be apparent if the syllables were extracted by the experimenter before the noise was added. A second feature of both repetitions and prolongations, already mentioned, is that they usually occur in initial position in a syllable. If a dysfluent syllable starts an extract, the position where the dysfluency occurs is provided to the listener. As with duration, noise might degrade listeners' ability to detect syllable position information yet listeners would be able to use it if syllables were spliced out of their context and presented for judgement.

By selecting test syllables that occur in a fluent context, all test intervals can have the same duration and the position where the test syllable occurs can be varied to obviate these

problems. This proposal has some similarities with Ingham et al.'s (1993) procedure. There are, however, two differences between their procedure and the current one that need highlighting: a. Here, the speech is processed so apparently fluent syllables that appear within supralexical dysfluencies are excluded. b. All syllables in an interval are assessed from low noise recordings to see if they contain a lexical dysfluency prior to selection of test intervals. Intervals for noise tests are then selected so that they contain one and only one lexical dysfluency. In contrast, in Ingham et al.'s (1993) time interval procedure, intervals are selected at random. Consequently, in their procedure it is not known what dysfluencies the intervals contain and whether a dysfluency is completed within an interval. A by-product of the selection criteria applied in the current study is an indication of the extent to which intervals include multiple dysfluencies or have dysfluencies split between one or more test interval/s. This topic is returned to in the Discussion when consideration is given to whether time interval analysis provides a satisfactory measure of stuttering.

B. Effects of noise on dysfluency assessment

Noise can originate in recording equipment and in the equipment used when the recordings are assessed subsequently as well as in external noise present in the recording environment. The ways the noise floor, frequency response and response to transients of the equipment can lead to poor registration of the intensity-time profile of speech have been documented by Rosen and Howell (1981) when the speech is initially recorded, and by Scott and Howell (1992) when the sounds are reproduced for perceptual testing. The main source of room noise in clinics is extraneous sound. Some extraneous noises are relatively easy for the experimenter to control (e.g. clinical personnel speaking or corridor noises). Others, such as hum due to equipment in the building and traffic rumble, are not under direct experimental control. In addition, to these sound sources, rooms with hard walls, like those in clinics, are reverberant and this also degrades recording quality (Watkins, 1992).

Noise from all these sources would make judges' decisions variable. It would be expected that judges would be less accurate when deciding about dysfluencies when noise level is high than when it is low. Some of the variability in Kully and Boberg's (1988) study that was attributed to variation between judges would arise from these sources. The design of the study did not allow control of reproduction equipment as the judges who acted as subjects were sent tapes that they played on their own equipment. The tapes judged on the poorest quality equipment would be more variable than those judged on equipment with better signal-to-noise ratio. In this way, inter-equipment variability would appear as inter-judge variability. Generally speaking, without agreed recording standards, it is not possible to compare results obtained in different recording and test environments.

The preceding observations underline why care has to be taken when recording and reproducing speech for research purposes. However, scant attention has been given to the quality of recordings used in research publications aimed at improving methods to collect and assess samples of stuttered speech. The minimum requirements for control of recordings for this purpose are to employ a calibrated acoustic environment during the initial recording and the subsequent testing phases and to ensure comparable quality recording and reproduction equipment are used always: If the equipment used is referred to explicitly, manufacturers' catalogs can be consulted for specifications and, when different equipment is used in different parts of a study, those specifications can be compared. None of these requirements is met in the Ingham et al., (1993) study that uses different recording and reproduction equipment and where tests are performed in widely differing environments.

It is expected that repetitions and prolongations will be affected in different ways when noise is introduced onto recordings. Considering repetitions first, one salient property is how intensity fluctuates over time. Each successive attempt of the iterated sound is followed by

an interval of relative silence. As noted earlier, poor quality acoustic environments affect intensity-time profiles. Consequently, detection of repetitions would be affected. In prolongations, where a sound is sustained, energy in the formant regions lasts for longer and allows listeners more time to integrate this information through the background noise. Consequently, prolongations would be expected to be less affected by noise.

In the present study, sections of speech were selected on the basis of the responses of expert judges which started and ended with fluent syllables and contained either a fluent syllable or one and only one syllable that was either a repetition or a prolongation (Boehmler, 1959). The number of sections that it is possible to select according to these constraints is described. An important subsidiary issue about this analysis is that it shows that a high proportion of samples are heterogeneous with respect to their dysfluency composition if these constraints are not applied (for example, when intervals are selected at random). The selected sections of speech were judged with different amounts of room noise added. The type of syllable contained in the fluent context determined what response was appropriate. Listeners judged the interval type when different amounts of noise were added. The hypotheses tested are that noise will increase error rates of repetitions and prolongations and that decisions about repetitions will be affected by noise more than prolongations.

1. METHOD

A. Subjects

Twenty-nine children who stutter were recruited. Twenty-five were male and four female. They ranged in age from seven years to 12 years (mean age 10.8 years). They were recorded when they were being assessed for treatment at a London clinic. The children had been referred to the clinic where they and their parents were seen by speech pathologists specializing in developmental stuttering. All the children employed were admitted to intensive therapy as a result of the assessments. The samples of speech were taken as part of their assessment.

B. Recordings

The speech material used in the current experiment was unscripted monologue speech samples, a minimum of two minutes long. A range of topics of conversation was suggested to the child before he or she started the monologue. Examples of topics are events on the way to the clinic, peer and sibling relationships, hobbies and favorite TV programs. Choice of which of these topics was taken up was left to the child. These topics were suggested so that the child did not “dry up” during the recording (this was never a problem).

The speech was transduced with a Sennheiser K6 microphone positioned six inches in front of the speaker in direct line with the mouth. All speech was recorded on DAT tape. The recordings were made in an Amplisilence sound-attenuated booth. The noise floor during recording was determined with the apparatus set up as in the test recordings. The noise floor was recorded while the speaker remained silent. This was fed into an Onno-Soki dual-channel analyzer. Background noise was more than 80 dB down relative to peak speech value between 500Hz and 10kHz. Speech was transferred digitally to computer for further processing. The speech from the DAT tapes was down-sampled during transfer to 20 kHz.

C. Location of fluent syllables, part-word repetitions and prolongations by expert panel

Expert judges were employed to assess the speech for selection of test material (details about the judges are given in section C.1). These judges provided two assessments of the speech: They located supralexical dysfluencies (section C.2). Speech was then syllabified

and all syllables classified as being fluent or containing designated types of lexical dysfluencies (section C.3).

1. Expert judges—Three expert judges were employed in this part of the experiment. Two of these were the judges employed in Howell et al. (1997b). These judges are both male and have considerable experience in categorizing stuttered events (seven years and two years respectively each with two years experience using procedures similar to here). They are researchers involved full time in developing assessment techniques and assessing linguistic factors in stuttering. The more experienced judge has phonetic training up to masters level and the other a doctorate in computational linguistics. The performance of these judges in tasks like here has been reported elsewhere (Howell et al. 1997b). The other judge is female with phonetic training up to masters level. She has one year's experience with stuttered speech, using similar procedures to the current ones. Once she was trained, inter-judge agreement was assessed by comparison with the other two judges. One of the expert judges from Howell et al. (1997b) only performed classification of syllables as lexical dysfluencies.

2. Location of supralelexical dysfluencies—The speech was transcribed as the first step in location of supralelexical dysfluencies. Two of the three judges transcribed the recordings independently. They did this by repeatedly listening to sections of the recording segmented into tone units. The transcription was in orthographic form with word attempts not indicated, e.g. k..k..Katy would be Katy. The agreement between transcribers was high (92% of all words) and so one transcriber's version was chosen at random for use in the subsequent assessments.

The two principal judges went through the transcriptions and located occurrences of phrase repetitions (e.g., “in the, in the morning”), phrase revisions (e.g., “my brother, er, no my uncle”), multi-word interjections (e.g., “you know”) and idea abandonments where one topic was abandoned and another commenced (see Wingate 1988 for detailed descriptions of these type of dysfluency). The extent of these supralelexical dysfluencies was determined based on Levelt's (1983) work examining similar structures to these that occur in fluent speakers' speech (Howell et al., 1997). Levelt identified several speech components round a dysfluency all of which occur in the phrase revision “You <turn left at the, no I mean, turn right at the> crossroads”. The word “left” is in error (the reparandum) which is later substituted by the word “right” (the alteration). The speaker has an overshoot (the words “at the” after the reparandum). Before the speaker recommences, an interjection (“no I mean”) is produced. When making the correction, the speaker backtracks (i.e. says the word “turn” a second time). Retraces, overshoots and interjections (multi- or single-word), all of which appear in this example, are optional parts of phrase revisions. Identification of these components allows speech to be parsed so supralelexical dysfluencies can be located and their extent determined. In the example, the supralelexical dysfluency involves all the words within <>. The extent of a phrase repetition was determined in like manner (the only difference between a phrase revision and a phrase repetition is that the repetition does not contain altered words). Incomplete phrases were identified as semantic discontinuities. Any sequence of words that was incomplete semantically was marked as an incomplete phrase and these words were excluded from interval selection.

3. Location of lexical dysfluencies—The speech was syllabified to select the stretches of sound to be presented to judges for assessment of lexical dysfluencies. All the speech was syllabified (including syllables that occurred within supralelexical dysfluencies). Syllables that were part of supralelexical dysfluencies were excluded before statistical analysis for lexical dysfluency and for selection of test material. Syllabification is described followed by the procedure adopted for assessing lexical dysfluency.

Syllabification: Syllabification of speech was done by an artificial neural network. The network was used as it allowed the large amount of speech available to be processed quickly and reliably. The syllabification algorithm first located the vowels in the speech. The architecture for vowel detection consists of a two-layer recurrent Elman network with 13 inputs, eight hidden units and one output unit, with back-propagation as the learning algorithm. The units in the input layer are fully connected to the units in the hidden layer (all of which have a single context unit) which are in turn fully connected to the output unit. The inputs to the network were the first 12 Mel Cepstral coefficients (10ms frames, 10ms step) plus one amplitude envelope parameter per frame. The envelope parameter was obtained by bandpass filtering the speech between 100 and 400 Hz using fourth-order Butterworth filters. The filter output was rectified and the resultant signal then smoothed with a 25 Hz low-pass, second-order Butterworth filter. The amplitude envelope was then summed over each 10 ms window and the base 10 logarithm of this taken. Each of the 13 input parameters was normalised to between -1 and +1 by subtracting the mean and dividing by twice the standard deviation. The networks were trained with the read speech from six male child stutterers reading “Arthur the Rat” that was used in Howell et al. (1997a). The children were aged between eight and 12 years. The first 30 seconds of speech from each speaker was used for training. During training, the network had to activate the vowel output once on each vowel and to remain inactive during other phonemes. The peaks associated with vowels were obtained by smoothing the network output with a 5 Hz low-pass, second-order Butterworth filter.

The syllable boundaries were located by using the vowel markers located by the networks together with a speech-silence detector (Rabiner & Sambur, 1975). If the Rabiner and Sambur (1975) algorithm detected silence between a pair of vowels the offset of the first syllable was placed at the start of the silence and the onset of the second syllable at the end of the silence. If there was no silence between a pair of vowels a single boundary was placed at the point of minimum energy between the pair of peaks. A voiceless repetition can be incorporated with the following syllable. After training the ANNs were tested on spontaneous speech from child stutterers. The networks correctly classified 95.6% of syllables in the spontaneous speech.

Assessment procedure: The procedure adopted for deciding about lexical dysfluencies by the experienced judges was developed from that employed by Howell et al. (1997b). Three independent assessments were made by each judge: In the first assessment, judges only indicated which syllables were fluent. Syllables were only considered fluent if all three judges labelled them fluent. These were not involved in subsequent assessments. For the remaining syllables there was some indication that the syllable was dysfluent. All the syllables that remained were assessed independently for category of dysfluency - once for repetition and once for prolongation. Note that since the second and third assessments were made independently and on the same material, a judge could give a prolongation response on one occasion and a repetition on another to a syllable (as would be appropriate in an example like “mmmm.m.mother”).

Other than the materials judged and the response allowed, the way the assessments were performed was common to all three of the assessments. The pool of syllables for making the judgement was specified (all syllables for fluent judgements or the dysfluent syllables for repetition and prolongation judgements). A random presentation order for the syllables in each pool was then computed so that the global context in which judgments were made was as constant as possible. The first randomly selected syllable and the syllable that followed were played to the judge. The pair of syllables had the same timing as they had had in the original recording so pauses were apparent as quiet sections between the two syllables in the pair. The response the judge made was whether the first syllable was fluent, repetition or

prolongation depending which assessment was being performedⁱ. For fluent and prolonged syllables, the second syllable simply served to provide context. The second syllable also provided some context for deciding about repetitions. It is not possible to determine whether a repetition that is split between two syllables (because the initial attempt has a vowel) is occurring unless the adjacent syllable is heard. So, for example, “cat” has to be heard to ascertain that the first syllable in “cuh-cat” is repeated. Location of repetition across pairs of syllables allows more extensive iterations to be identified. For instance, in a sequence like “cuh-cuh-cat”, repetition would be detected on both the first and final pair of syllables. Though consecutive runs of repetition like this can be designated as iterations of a single repetition, they are long and tend to straddle test intervals (see 1.4 in section D below). The syllable pair being tested could be heard repeatedly (by pressing the return key on a keyboard) until the judge was ready to make his or her response. The judges were self paced and could take a break whenever they wanted. Test sounds were played over RS 250-924 headsets in the test booth. To check on noise level added to the signal from external and electrical sources, a speech signal was input at 75 dB SPL. Background noise with speech switched off was 40 dB SPL below the speech level.

D. Interval selection criteria

Once the syllables had been assessed, they were used to select test materials. Test material had to have a designated target syllable and fluent syllables around it. Moreover there had to be sufficient fluent syllable context to allow a range of starting positions to be used for the target syllable while at the same time overall duration had to be kept constant for the reasons given in the introduction. These constraints required that the initial fluent context extended beyond the test interval duration. Intervals of the same duration where the target syllable has different starting positions can then be selected from the more extensive context. These more extensive contexts are referred to as super-intervals for brevity. Test intervals are the parts of super-intervals presented for assessment. In the following sub-sections, selection criteria for target syllables and test intervals are described followed by the procedure for adding noise.

1. Target syllable—Each syllable was checked to establish whether it met the following criteria for designation as fluent, repetition or prolongation:

- 1.1. A syllable was designated fluent if all judges gave it a fluent response.
- 1.2. For a repetition syllable, at least two of the judges had to call it a repetition. If only two judges considered the syllable to be a repetition, neither of these judges was allowed to give this syllable a prolongation response as well (this excluded dysfluencies like the “mmmm.m.mother” example in section C.3 above). The response of the judge who did not designate the syllable as a repetition was not allowed to be “prolongation”.
- 1.3. The criteria for a prolongation syllable were the same as those for a repetition syllable given in 1.2 except that the dysfluency type specifications were reversed.
- 1.4. An additional criterion applied solely to repetitions. They could be part- or whole-syllable but they could contain only one repeated syllable (eg “ka Katy” not “ka ka ka Katy”). This restriction was applied so that repetitions were limited in extent to a single interval and so that the syllables in intervals of different target syllable types could be found which corresponded roughly in duration (see 3.1). Though this criterion could

ⁱHowell et al. (1997b) took ratings as well as categorisations of each word. Here, since a panel of judges is used, it is possible to dispense with ratings: Control experiments have shown that the results of the panel are monotonically related to ratings given to each syllable.

conceivably have been required for long prolongations, this was not necessary for the present speech material.

1.5. The “target” syllable was not allowed to occur within any part of a supralelexical dysfluency.

Laxer criteria were adopted in 1.2 and 1.3 because there were fewer syllables in repetition and prolongation classes than there were fluent syllables.

2. Interval construction criteria—Once target syllables that had a fluent, repetition or prolongation designation had been obtained, they were checked to see whether they could be used to construct a super-interval. All appropriate target syllables were checked. The principal requirement behind selection of super intervals was to provide fluent surrounding context for the designated target syllable. The criteria for the super-intervals were:

- 2.1. A super-interval could contain one and only one syllable that was fluent, repetition or prolongation and no syllable that was part of a supralelexical dysfluency. All other syllables had to be agreed by all judges as fluent syllables.
- 2.2. Isolated interjections (including filled pauses such as “umm” or “err” could not occur within any of the super-intervals.
- 2.3. There had to be two unanimously agreed fluent syllables preceding the target syllable.
- 2.4. The target syllable had to have three unanimously agreed fluent syllables following it.

If the fluent fillers in super-intervals had a fluent context word which a majority but not all judges designated fluent (only two rather than three judges passing constraint 2.1 on one fluent context word), but passed all other selection criteria, they were retained as potential practice material for the test judges.

A super-interval might consist of the words “the black mmmouse sat on the” where “mmmouse” is a prolonged target syllable. From this super-interval, three test intervals were constructed. The first of these started at the first syllable in the super-interval and extended to the end of the fluent syllable after the prolonged syllable (in the example, the interval “the black mmmouse sat”). This interval could include silent pauses. The duration was measured and the subsequent two test intervals had to be close in duration to it. The second interval started on the second syllable (on “black” in the example) and extended at least one fluent syllable beyond that which ended the first test interval. The third test interval started on the target syllable itself and continued to at least one fluent syllable beyond the end of the second test interval.

3. Selection of fluent and repetition super-intervals to match prolongation super-intervals—After the restrictions outlined in sections D.2 and D.3 above, only ten prolongation super-intervals were available (with the three starting positions used in each super-interval, this allowed construction of 30 prolongation test intervals). Equal numbers of fluent and repetition super-intervals were chosen. The fluent and repetition super-intervals were selected so that they shared the same basic temporal features as the available prolongation intervals. The selection criteria for the repetition and prolongation super-intervals were:

- 3.1. The syllables in the fluent, repetition and prolongation intervals had to have similar average durations.

3.2. Pauses, if they appeared, had to be in corresponding positions. Unlike the previous criteria, these were applied manually. A breakdown about how the main selection criteria whittled down the instances is presented in the results.

4. Addition of noise to test intervals—The ten super-intervals of each of the three target syllable types allowed 90 test intervals to be constructed when the three starting positions were applied. Three further stimuli were created from each of these 90 stimuli. The first had no noise added, the second 3 dB and the third 6 dB of noise relative to the mean speech amplitude of each syllable.

Noise was real room noise recorded on DAT tape with the Sennheiser K6 microphone. The noise was recorded in a quiet office that had plaster walls with nothing on them and was 8 feet $\times$ 20 feet. It had a window at one of the narrow ends that was closed and a door at the opposite end that was also closed while the recording was made. The office was located in the center of London and there was background traffic noise. The recording equipment was in the room and it was switched on. The equipment and microphone were located in the center of the room. A section of the recorded noise was selected where there was no noise from the corridor and where there was no predominant sound such as traffic accelerating. The noise was typical of that which occurs in the best recordings made in similar rooms in clinics. To check this, the signal to noise (S/N) ratio was calculated on four recordings supplied by US clinics using the algorithm described by Sims (1985). Noise levels over 40 second extracts were between 2.7 dB and 6.7 dB.

The processing to add the noise to the speech samples was as follows:

- 4.1. The test intervals were scaled to the same maximum value.
- 4.2. The mean amplitude was calculated over the full length of the test interval.
- 4.3. Each test interval was processed separately for each desired noise level. S/N ratio was equal to $10 \times \log_{10}$ (mean amplitude of S/mean amplitude of N).

The mean amplitude of N was calculated over the length of the interval to be processed (i.e. if the interval was 1,200 samples long, the mean amplitude of the noise was calculated over the first 1,200 samples and N for this interval stopped at this sample). A multiplier was applied to the noise to bring the desired mean amplitude of N to the value appropriate to give the required S/N ratio. The S and N were then added.

E. Experimental procedure for assessment of test intervals by test judges

Six test judges were employed for the experiment. These were all female students who were in their fourth (final) year of a full-time speech pathology course. They had received training on clinical methods of assessing dysfluency as part of their course and had experience of making such judgements in clinical practice. They were chosen as they had a homogeneous level of experience and were representative of the clinical judges who would be required to assess clients' recordings. These were used to performing psychometric assessments. The playback apparatus for the test judges was the same as that described earlier. The 270 test intervals (the 90 test intervals described at the beginning of the previous section $\times$ three for intervals at different noise levels) were selected at random without replacement and these were presented to the test judges to make their response. The response was whether the interval contained only fluent syllables or included either a repetition or a prolongation. Only one response was allowed to each interval heard. Each judge did the experiment in isolation. They pressed a mouse button and they heard each stimulus once over the RS 250-924 headset.

The subjects received practice before they did the assessments proper. Only three prolongations met all the selection criteria except that one of the fluent context syllables was not unanimously agreed to be fluent (criterion 2.1) during the assessment by the expert judges. Three fluent and repetition super-intervals (matched in the same way as the test super-intervals) were selected. Eighty one practice test intervals were created from the original three fluent, prolongation and repetition super-intervals. These were tested prior to collection of data on the experimental material in the same manner. Responses to the practice intervals were discarded before statistical analysis.

II. RESULTS

A. Effect on distribution of intervals of selection criteria

The way the selection criteria restricted the candidate super-intervals was examined. Each syllable was considered with respect to whether there was agreement across judges about whether the syllable was fluent, a repetition or a prolongation according to criteria D.1.1 to D.1.4 specified in the method. Syllables that met the agreement criteria were potential candidates for target syllables and, in the case of agreed fluent syllables, as filler items for the superinterval. Over all syllables, the proportion that did not meet these criteria was 18.2% (1,699 syllables). Note that not all the syllables that failed to meet the agreement criteria are disagreements between judges. This 18.2% includes cases where the word was considered dysfluent by all judges but none of the judges considered it a repetition or prolongation and also cases where two or three of the judges agreed that a target syllable was both repetition and prolongation as in the “mmmm.m.mother” example given in C.3. The number of agreed syllables is given separately for fluent, repetitions and prolongations in row one of Table 1.

The second row of Table 1 shows the number and percentages of super-intervals removed because the super-interval had a word that was part of a supralelexical dysfluency. Exclusions are classified by target syllable type even though the super-interval could be prohibited because the target syllable (criterion D.1.5) and/or syllables in the super interval filler context (criterion D.2.1) occurred within a supralelexical dysfluency. Also note that these data do not include standalone interjections. The average over all target syllable types of super-intervals where the syllables in an interval overlapped with a supralelexical dysfluency was roughly 40%. A further factor of note about the number of target syllables that occurred within a supralelexical dysfluency was that the probability that a repetition will be involved in a supralelexical dysfluency was higher than that for a prolongation. A Chi square on the number of repetitions and prolongations that were part of a supralelexical dysfluency (459 and 29 respectively, as shown in row two) against those not part of a supralelexical dysfluency (the sum of the remaining columns in Table 1, 337 and 57 for repetitions and prolongations respectively) showed that there was a significant association (Chi square = 18.0, df = 1, $p < 0.001$)

The remaining analyses summarised in Table 1 examined the syllables around the target syllable to see whether they fulfilled the criteria for specification of a super-interval with the response designation of the target syllable. The third row of Table 1 shows the reduction in number of super-intervals caused by the occurrence of interjections that were not part of supralelexical dysfluencies (D.2.2). From the percentages of super-intervals that contained such an isolated interjection, it appears that these depend on target syllable type (15.3, 9.5 and 19.8% for fluent, repetition and prolongation respectively). This would support the view that there is relative inhibition of interjections around repetitions and a heightened chance around prolongations relative to fluent contexts. However, summing up the syllables in the remaining rows for the respective columns to estimate the number of these that did not contain an isolated interjection gives totals of 2,812, 261 and 40 compared with 1032, 76

and 17 that contain an interjection. A Chi square test showed that there was no association between target syllable type and whether the super-interval contained an interjection (Chi square = 3.25, $df=2$, $p>0.05$). This is particularly noteworthy bearing in mind the high number of fluent target syllables that would tend to lead to a significant result. The same Chi square analysis using repetition and prolongation alone was also not significant (Chi square = 1.43, $df=1$, $p>0.05$). The relative reduction in the percentage of super-intervals constructed from repetition target syllables that contained an interjection was due to the higher likelihood of repetitions being associated with supralelexical dysfluencies that reduced the overall total. Thus, the likelihood of an isolated interjection occurring in a super-interval was about the same whether the target syllable was fluent, repetition or prolongation.

Row four shows the reduction in super-intervals brought about because the interval contained a lexical dysfluency on a syllable other than the target syllable or because the super-interval contained a syllable other than the target syllable that was not agreed fluent (D.2.3 and D.2.4). A similar analysis to that on isolated interjections was done on the lexical dysfluencies. Again the percentages suggest that prolongations are hit harder by lexical dysfluency prohibitions than are repetitions. The totals in the last two rows of Table 1 were entered into a contingency table where one row was whether the lexical dysfluency criteria prohibited the candidate super interval being accepted (626, 192 and 30 for the fluent, repetition and prolongation columns respectively) and row two was not prohibited by lexical dysfluencies (2,186, 69 and 10 for fluent, repetition and prolongation). Chi square analyses showed that there was a significant association when fluent syllables were included (Chi square = 363.77, $df=2$, $p<0.001$) but not when repetitions and prolongations alone were considered (Chi square = 0.037, $df=1$, $p>0.05$). Reservations about the Chi square analysis with fluent syllables need to be expressed given the sensitivity of this statistic to high numbers of observations that can lead to spurious significance as mentioned earlier. It would be more conservative, therefore, to conclude that there was no association between whether an interval contained a prohibited lexical dysfluency and repetition and prolongation dysfluency target syllable type. In summary, it appears, then, that the only factor leading to super-intervals being rejected in Table 1 that differentially affected target syllable types was the high proportion of repetitions that occurred as part of a supralelexical dysfluency.

A further factor of note is that the number of super-intervals was dramatically reduced by the specified criteria (bearing in mind that over an hour's speech, in total, was assessed). A response to an interval as stuttered or non-stuttered is only meaningful if the interval contains a specific type of dysfluency and the interval incorporates all the dysfluency. The other side of this analysis is, therefore, that it shows arbitrarily chosen intervals are not appropriate for assessing dysfluency type: If this is done, many intervals will contain multiple dysfluencies (lexical and supralelexical) and syllables that expert judges do not agree about. The selection criteria also prevent a dysfluency from being split between different intervals (fragmentation). Minimizing the occurrence of multiple dysfluencies in an interval and prevention of fragmentation of dysfluencies between intervals make incompatible demands about choice of interval duration: Multiple dysfluencies can be reduced by choosing shorter intervals whereas fragmentation is reduced if intervals are made longer. Despite the restrictions on choice of super-intervals, some intervals were found that contained a target syllable of the designated type that occurred in the context of fluent syllables. The fluent syllable contexts, in turn, allowed test intervals to be constructed to establish how noise and starting position affected detectability of dysfluencies by test judges. The selection constraints also determined the duration of the super-intervals. The duration of super-intervals that passed all tests was 2.71s for fluent intervals, 2.64s for prolongation intervals, and 2.60s for repetition intervals.

B. Effects of noise on detectability of repetitions and prolongations

Error proportions and standard deviations for each target syllable type at the three starting positions and when the three levels of noise were added are shown in Table 2 averaged over judges. Error rates were relatively high even when there was no noise added to the test intervals because the test judges were only allowed to hear the sound once (when they signalled that they were ready). These data were analyzed by a three-way analysis of variance (ANOVA) with factors starting position relative to the super-interval (three levels; two, one or no fluent syllable lead in), noise level (three levels; no noise, 3 dB and 6 dB) and target syllable type (three levels; fluent, repetition or prolongation targets).

The main effect of noise level was significant ($F_{2,135} = 26.18$, $p < 0.001$). This arose, as expected from higher error rates as noise level increased. Tukey simultaneous tests showed that error rate was significantly higher in the 3 dB ($T = 3.9$, $p < 0.001$) and 6 dB conditions ($T = 7.2$, $p < 0.0001$) after adjustment for error rate in the no noise condition and error rate was significantly higher in the 6 dB condition ($T = 3.3$, $p < 0.005$) after adjustment for error rate in the 3 dB condition. The main effect of target syllable type was also significant ($F_{2,135} = 123.39$, $p < 0.001$) with more errors on repetitions than prolongations and more errors of prolongations than fluents. Tukey simultaneous tests showed that error rate was significantly higher for repetitions ($T = 8.8$, $p < 0.0001$) and prolongations ($T = 15.6$, $p < 0.0001$) after adjustment for fluent error rate and error rate for prolongations was significantly higher ($T = 6.9$, $p < 0.0001$) after adjustment for error rate on repetitions.

The hypothesis, presented in the introduction, that repetitions will be more affected than prolongations by noise predicts an interaction between noise level and target syllable type. This interaction was significant ($F_{4,135} = 5.17$, $p = 0.001$). Inspection of the data in Table 2 shows that the percentage increase in errors with increasing noise level was biggest for repetitions for each of the three starting positions (the difference in percentages between no noise and 6 dB of noise was 38.3, 33.3 and 31.6%), next biggest for prolongations (20, 15.0 and 15.0%) and least of all for fluents (5.0, 10.0, 3.4%). Thus, it appears to be the case that repetitions are most affected by noise and prolongations next-most affected, as hypothesized. The interaction between starting position and target syllable type was also significant ($F_{4,135} = 2.76$, $p = 0.03$). The data show that prolongations are least error prone when they appear late in the test interval (a bigger preview), whereas the converse is true of repetitions and fluents.

III. DISCUSSION

A. Effect on distribution of intervals of selection criteria

The findings have relevance for choice of an appropriate assessment procedure for research purposes on stuttering in general. The current study shows that intervals less than three seconds long that contain one and only one type of lexical dysfluency are rare in connected speech. More intervals would occur if longer lexical and suprallexical dysfluencies (occupying up to the entire interval) were allowed. Such long dysfluencies are more likely to be fragmented between intervals when the intervals are randomly imposed (this can also happen even with short lexical dysfluencies). This can lead to one dysfluency affecting two adjacent intervals. It can also lead to dysfluencies being missed as when word or phrase repetitions consisting of a single repetition are split at the point where the repetition recommences. These observations make application of time-interval analysis procedures to, for instance, assessing the effects of frequency-shifted feedback (Ingham, Moglia, Frank, Ingham, & Cordes, 1997) problematic.

Time interval analysis procedures have the advantage that they are relatively quick and easy to perform in comparison with current assessments. It is not clear, however, how the results

of such procedures should be interpreted. Though they appear to give an indication of treatment outcome, as discussed in the preceding paragraph (Ingham et al., 1997) they bias these estimates by underestimating dysfluencies on some occasions and inflating them on others. The dominant effect appears to be to overestimate stuttering. Thus, as seen in Ingham et al.'s (1997) data where 5-second randomly-selected intervals were used, the vast majority of intervals are judged as dysfluent leading to ceiling effects. The problem in use of this technique for diagnostic and prognostic indications is that it does not allow dysfluency types, known to be important for each of these topics, to be assessed. These indicators are discussed further in the following section.

B. Effects of noise on detectability of repetitions and prolongations

The effects of noise on assessments of repetition and prolongation show that comparison of performance across recording environments with different noise levels distorts fluency assessment. They also underline the fact that details of the recording environment (including microphones, test rooms and so on) need to be as noise-free as possible. Without these precautions, it is likely that stuttering will be underestimated since the results show that repetitions, in particular, as well as prolongations to some extent tend to be designated fluent whereas fluent intervals are relatively infrequently called repetition or prolongation. There should also be agreed specifications about minimum recording standards that need to be met to conform to clinical and research requirements.

The first specific implication about noise affecting dysfluent classes differentially follows from the finding that noise leads to stuttering incidence being underestimated. This has significance for diagnosis as missed dysfluencies, by definition, would underestimate stuttering frequency. Similarly, if by chance noise level increases over recordings made during and after treatment, this could lead to spurious apparent improvements. It would also lead to prognostic indications being missed and distorted. So, for instance, Conture's (1990) view is that a change from a predominance of repetitions in a child's speech to a high proportion of prolongations is a sign that the disorder is worsening. This would appear to happen if speech recorded from a noisy environment was compared with one recorded in a quiet environment due to noise affecting repetitions more than prolongations. Alternatively if an initial assessment made in quiet was compared with one in noise, a judge might miss this telltale sign.

A second important finding in this study is that detectability of repetitions and prolongations in noise is affected by the position that they occupy in a test interval. For one thing, if intervals are used, the differential detectability could affect diagnostic and prognostic indications in a similar way to that discussed in the previous paragraph. Also, the findings may have theoretical significance. A breakdown in fluency that results in a prolongation may arise from speech changes made in the prior context whereas a repetition breakdown may occur before a problematic word (Au-Yeung et al., in press).

One difference between the stimuli used in Ingham et al.'s (1997) time interval analysis procedure and the current is that the former authors used audiovisual rather than audio alone stimuli. This might be considered a limitation to the investigation of noise in the present study since audio-visual presentation is widely regarded as allowing judges to be more sensitive to stuttering than audio alone. However, this matter is not at all clear cut. Judgements of audiovisual synchrony are less sensitive when in the familiar task of viewing a face when speaking than the less familiar task of watching a person hammering (Dixon & Spitz, 1980). The well-known McGurk effect can also be interpreted as showing a lack of sensitivity to audio information when a face is viewed (McGurk & MacDonald, 1976). In this illusion, the presence of a facial view of a speaker producing a /b/ dubbed with the speaker's production of /g/ leads to perception of /d/. The McGurk effect appears to make a listener

less sensitive to the auditory features responsible for perception of plosive stop consonants. A second limitation is that we have not examined the extent to which reduction in intervals due to the selection criteria are ameliorated by allowing more extensive dysfluencies that still occur within the confines of an interval of a specified duration. Though this would make the problem somewhat less acute, it would not remove the problem.

Acknowledgments

This research was supported by a grant from the Wellcome Trust.

References

- Alfonso P. Subject definition and selection criteria for stuttering research in adult subjects. *ASHA Reports*. 1990; 18:15–24.
- Au-Yeung J, Howell P, Pilgrim L. Phonological Words and Stuttering on Function Words. *Journal of Speech, Language and Hearing Research*. in press.
- Boehmler RM. Listener responses to non-fluencies. *Journal of Speech and Hearing Research*. 1959; 1:132–141. [PubMed: 13576506]
- Brown SF. The loci of stuttering in the speech sequence. *Journal of Speech Disorders*. 1945; 10:181–192.
- Conture, EG. *Stuttering*. 2nd Edition. Englewood Cliffs, New Jersey: Prentice-Hall; 1990.
- Dixon NF, Spitz L. The detection of auditory visual desynchrony. *Perception*. 1980; 9:719–721. [PubMed: 7220244]
- Howell P. Stuttering in childhood. *The Ciba foundation Bulletin*. 1993; (35):15–16.
- Howell P, Au-Yeung J, Sackin S, Glenn K, Rustin L. Detection of supralexical dysfluencies in a text read by child stutterers. *Journal of Fluency Disorders*. 1997; 22:299–307. [PubMed: 18259593]
- 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 Recognition of Repetitions and Prolongations with Supplied Word Segment Markers. *Journal of Speech, Language and Hearing Research*. 1997a; 40:1085–1096.
- 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: I. Psychometric Procedures Appropriate for Selection of Training Material for Lexical Dysfluency Classifiers. *Journal of Speech, Language and Hearing Research*. 1997b; 40:1073–1084.
- Howell, P.; Sackin, S.; Glenn, K.; Au-Yeung, J. Automatic stuttering frequency counts. In: Peters, H.; Hulstijn, W.; Lieshout, P., editors. *Speech Motor Production and Fluency Disorders*. Amsterdam: Elsevier; 1997.
- Howell P, Young K. The use of prosody in highlighting alteration in repairs from unrestricted speech. *Quarterly Journal of Experimental Psychology*. 1991; 43(A):733–758. [PubMed: 1775664]
- Ingham RJ, Cordes AK, Gow ML. Time-interval measurement of stuttering: Modifying interjudge agreement. *Journal of Speech and Hearing Research*. 1993; 36:503–515. [PubMed: 8331907]
- Ingham RJ, Moglia RA, Frank P, Ingham JC, Cordes AK. Experimental investigation of the effects of frequency-altered auditory feedback on the speech of adults who stutter. *Journal of Speech, Language and Hearing Research*. 1997; 40:361–372.
- Kully D, Boberg E. An investigation of inter-clinic agreement in the identification of fluent and stuttered syllables. *Journal of Fluency Disorders*. 1988; 13:309–318.
- Levelt WJM. Monitoring and self-repair in speech. *Cognition*. 1983; 14:41–104. [PubMed: 6685011]
- McGurk H, MacDonald J. Hearing lips and seeing voices. *Nature*. 1976; 264:746–748. [PubMed: 1012311]
- Parducci A. Category judgment: A range-frequency model. *Psychological Review*. 1965; 17:9–16.
- Rabiner LR, Sambur MR. An algorithm for detecting the endpoints of isolated utterances. *Bell Systems Technical Journal*. 1975; 54:297–315.

- Rosen S, Howell P. Plucks and bows are not categorically perceived. *Perception and Psychophysics*. 1981; 30:1256–1260.
- Scott, S.; Howell, P. Infinitely peak clipping speech alters its P-center; *Proceedings of the Fourth International Congress on Rhythm*; 1992. p. 151-156.
- Sims JT. A speech-to-noise ratio measurement algorithm. *Journal of the Acoustical Society of America*. 1985; 78:1671–1674. [PubMed: 4067080]
- Smith, A. A multilayered dynamic approach to stuttering. In: Peters, H.; Hulstijn, W.; Lieshout, P., editors. *Speech Motor Production and Fluency Disorders*. Amsterdam: Elsevier; 1997.
- Watkins AJ. Perceptual compensation for the effects of reverberation on amplitude envelopes - cues to the slay-splay distinction. *Proceedings of the Institute of Acoustics*. 1992; 14:125–132.
- Wingate, ME. *The structure of stuttering: A psycholinguistic study*. New York: Springer-Verlag; 1988.

Table 1

Breakdown of how super-intervals (determined by the type of target syllable shown along the top row) are reduced by the selection requirements indicated in the left column. The totals and percentages (in brackets) are given.

Target syllable type	Fluent	Repetition	Prolongation
Overall target syllables of the designated type	6767	796	86
One of the syllables in a super-interval occurs within an SD.	1718 (25.4%)	459 (57.6%)	29 (33.7%)
Super-interval contains an interjection.	1032 (15.3%)	76 (9.5%)	17 (19.8%)
Super-interval contains another LD or a non-agreed F, P or R syllable.	626 (27.1%)	192 (24.1%)	30 (34.9%)
Remainder	2186 (32.3%)	69 (8.7%)	10 (11.6%)

Table 2

Proportion of errors and their standard deviations (in parentheses beneath) for test intervals. Target syllable type is given in the column on the left (labelled F for fluent, R for repetition and P for prolongation). The three starting positions relative to the super-interval are indicated along the top row at the head and the three different noise levels are labelled in the following row.

	Start 1			Start 2			Start 3		
	No noise	3dB	6dB	No noise	3dB	6dB	No noise	3dB	6dB
F	13.3 (12.1)	11.7 (16.0)	18.3 (14.7)	3.3 (5.1)	8.3 (9.8)	13.3 (15.1)	8.3 (9.8)	6.7 (8.2)	11.7 (16.0)
P	28.3 (14.7)	36.7 (13.7)	48.3 (16.0)	28.3 (11.7)	35.0 (8.4)	43.3 (12.1)	18.3 (11.7)	28.3 (17.2)	33.3 (19.7)
R	26.7 (17.5)	50 (16.7)	65 (11.7)	35 (15.2)	58.3 (14.7)	68.3 (9.8)	36.7 (12.1)	55.0 (12.2)	68.3 (9.8)