APHASIOLOGY, 1998, VOL. 12, NO. 7/8, 715–730

Single word intelligibility in aphasia and apraxia of speech

KATARINA L. HALEY†, ROBERT T. WERTZ‡§ and RALPH N. OHDE§

† Division of Speech and Hearing Sciences, Department of Medical Allied Health Professions, University of North Carolina, Chapel Hill, NC, USA
‡ Nashville Veterans Administration Medical Center, USA
§ Department of Hearing and Speech Sciences, Vanderbilt Bill Wilkerson Center for Otolaryngology and Communication Sciences, Nashville, TN, USA

Abstract

Single word speech intelligibility was evaluated in three groups: aphasia with apraxia of speech, aphasia with no apraxia of speech, and normal controls. Intelligibility was significantly lower in the two aphasic groups compared with the normal group and intelligibility did not differ significantly between the aphasia and apraxia of speech and the aphasia only groups. Seventy per cent of the speakers with apraxia of speech obtained intelligibility scores below the normal range and 80% of the speakers with aphasia only obtained intelligibility scores within the normal range. There was a moderate, but statistically non-significant, correlation between intelligibility and severity of apraxia of speech on an eight-point rating scale.

Introduction

Speech intelligibility, or the degree to which a speaker's utterances are understood by listeners, has become an important concept in speech–language pathology. Considerable attention has been given to the assessment of speech intelligibility in a variety of conditions that affect a person's speech production abilities, including hearing impairment (Osberger 1992), developmental phonological and articulation disorders (Kent *et al.* 1994), and dysarthria (Yorkston *et al.* 1992). Two basic strategies for quantifying intelligibility are available. The first is to have listeners make an overall estimate of how well a speaker can be understood and then assign this estimate to a level on a numerical rating scale. Another approach is to obtain a score experimentally by presenting a number of words or sentences produced by a speaker to a panel of listeners, having the listeners indicate what they think the intended utterances were, and then calculating the proportion of utterances that were understood. In comparison, the experimental approach is the most objectively defined and probably the most reliable of the two.

Although it is well recognized that speech produced by patients with apraxia of speech is perceived as containing numerous articulatory errors (Johns and Darley

Address correspondence to: Katarina Haley Ph.D., Division of Speech and Hearing Sciences, Department of Medical Allied Health Professions, Medical School Wing D, CB#7190, University of North Carolina, Chapel Hill, NC 27599–7190, USA.

0268–7038/98 $12·00

1970, Odell *et al.* 1990) and despite the widespread use of intelligibility measures in other speech production disorders, intelligibility testing has not typically been reported in the literature for patients with apraxia of speech. A few studies report the use of a rating scale (e.g. Seddoh *et al.* 1996). However, to our knowledge, there have been no systematic attempts to examine the intelligibility of speech produced by apraxic speakers. One of the greatest potentials for speech intelligibility assessment in apraxia of speech is as an objective, functional, and reliable measure of overall degree of speech involvement. Such a tool would be of substantial clinical importance in the management of patients with apraxia of speech as a change measure and as one form of severity indicator.

The present investigation was conducted to examine speech intelligibility in a clinical population of patients with coexisting aphasia and apraxia of speech (A-AOS).[1] We also wanted to compare the intelligibility of these individuals to that of aphasic subjects without apraxia of speech. There are a number of ways in which aphasia by itself may affect speech production. For example, word retrieval difficulties and phonemic as well as verbal paraphasic errors could all potentially compromise speech intelligibility. By including a comparison group of subjects with aphasia only, it was possible to address whether or not speech intelligibility is more compromised in aphasia with a coexisting apraxia of speech than in aphasia alone.

A number of issues must be considered when selecting a testing procedure for intelligibility. For example, decisions must be made regarding what type of speech material (e.g. single words vs. phrases), listening task (e.g. multiple choice vs. orthographic transcription), and listening context (e.g. presence vs. absence of semantic/syntactic cues) to use. Another important factor that requires attention is how to ensure experimental control, so that similar test conditions are provided for all speakers while, at the same time, listener learning effects are minimized. Because of the language problems of the populations of interest in this study, we elected to use single words instead of sentences. It was our intention thereby to limit the influence of potentially confounding problems with memory and syntax processing. In regard to word length, there are some potential advantages with multisyllabic words in that they are probably more likely to elicit incorrect productions than are monosyllabic words. Several clinical and transcription reports indicate that the frequency of errors in patients with A-AOS increases with increasing word length (Gandour *et al.* 1993, Johns and Darley 1970, Wertz *et al.*

[1] Among persons with apraxia of speech, there are a small number of cases that present with no or very minimal coexisting language involvement. The terms 'pure apraxia of speech' (Square *et al.* 1982) and 'aphemia' (McNeil *et al.* 1997, Pellat *et al.* 1991, Schiff *et al.* 1983) have been used to refer to patients with these symptoms. Careful examination and experimentation using subjects with pure apraxia of speech is of considerable theoretical interest. Such an approach has great potential for the development of models of speech production and for a better understanding of brain–behaviour relationship regarding higher level speech programming. However, the speech characteristics in patients with pure apraxia of speech are not necessarily the same as those of patients with A-AOS. In this latter group, speech production may be affected by motor impairment, linguistic impairment, or both. It is reasonable therefore to expect that the resulting speech in A-AOS may differ somewhat from that of pure apraxia of speech as well as from that of aphasia without apraxia of speech. Similarly, with regard to errors affecting individual speech segments, we must assume potential involvement at both linguistic and motor levels. We do not yet have a sufficient understanding of normal and abnormal speech production to allow a separation of 'apraxic' from 'phonemic paraphasic' errors in the speech of these patients with coexisting aphasia and apraxia of speech.

1984). Thus, error sensitivity may be higher in a multisyllabic than in a monosyllabic intelligibility test. On the other hand, multisyllabic words pose a greater challenge for experimental control than do monosyllabic words. For example, it is not possible to use the same set of multisyllabic words for all speakers due to listener learning effects. In contrast, one of the advantages with monosyllabic words is that they are easily confused with other monosyllabic words, particularly if minimally contrasting words are used. Thus, the risk that listeners will memorize the target words upon hearing repetitions of the same word list is reduced. A monosyllabic intelligibility test developed for dysarthria patients (Kent *et al.* 1989) served our purposes well. This test also has the advantage of including a broad range of phonetic contrasts, which makes it potentially sensitive to a variety of error patterns, as we might expect to find in A-AOS.

Thus, using the Kent *et al.* (1989) protocol, we compared single word intelligibility in speakers with A-AOS, speakers with aphasia only, and normal controls. Four questions were asked:

(1) Can intelligibility be tested reliably in patients with A-AOS?
(2) Is the speech in speakers with A-AOS significantly less intelligible than in speakers with aphasia only and in normal speakers?
(3) What is the relationship between single word intelligibility and aphasia severity?
(4) What is the relationship between single word intelligibility and overall apraxia of speech severity as estimated with a traditional, eight-point rating scale?

Methods

Subjects

Data were collected from three groups of subjects. There were 10 patients diagnosed with A-AOS, 10 patients diagnosed with aphasia without apraxia of speech, and 10 age-matched normal controls. All subjects were native speakers of American English. They had no structural vocal tract abnormalities that could potentially compromise articulation. Please refer to table 1 for a general summary of subject characteristics.

Subjects with aphasia and/or apraxia of speech were referred by speech–language pathologists and neurologists in the region. Patients who expressed an interest in participating were contacted by telephone or through an appointment scheduled in conjunction with their regular therapy. The aetiology was restricted to a cerebrovascular accident (CVA), as confirmed by a review of the patient's medical record. All patients were at least 3 months post-onset of the CVA at the time of the recording in order to ensure that they were relatively stable neurologically (Culton 1969). Aside from the speech and language sequelae of the CVA, none of the patients exhibited other symptoms of central or peripheral nervous system pathologies (e.g. dementia, dysphagia). Four of the patients (two in the aphasia group and two in the A-AOS group) had suffered more than one CVA. In these patients, time post-onset was calculated from the most recent event.

Patients who were referred to the study, and agreed to participate, completed speech and language testing in order to determine their appropriate diagnosis and to ensure that they met selection criteria. The aphasia quotient subtests of the

Table 1. Summary of subject data for the normal (NOR), aphasia and apraxia of speech (A-AOS), and aphasia only (APH) groups

Subject	Age	Gender	Educ.	Handedness	Dialect	MPO	Aphasia	AQ	Fluency	AOS	Dys.
NOR1	50	F	15	Right	Southern	NA	NA	99·8	10	0	0
NOR2	71	F	16	Right	General	NA	NA	99·9	10	0	0
NOR3	56	M	16	Right	General	NA	NA	99·8	10	0	0
NOR4	59	M	21	Right	General	NA	NA	100·0	10	0	0
NOR5	52	M	12	Right	Southern	NA	NA	99·4	10	0	0
NOR6	59	M	13	Right	Southern	NA	NA	97·0	10	0	0
NOR7	70	M	18	Right	Southern	NA	NA	100·0	10	0	0
NOR8	69	F	12	Right	Southern	NA	NA	99·3	10	0	0
NOR9	72	M	18	Both	General	NA	NA	98·8	10	0	0
NOR10	63	F	16	Right	General	NA	NA	100·0	10	0	0
Mean	62		16					99·4	10	0	0
A-AOS1	57	M	14	Right	Southern	45	Broca	59·5	4	5	2
A-AOS2	64	M	12	Right	Southern	11	Broca	73·9	4	4	1
A-AOS3	52	F	14	Right	Southern	49	Broca	61·2	4	2	0
A-AOS4	45	F	14	Left	General	30	Broca	72·3	4	4	0
A-AOS5	47	M	14	Right	General	63	Broca	67·5	4	1	0
A-AOS6	57	M	11	Both	General	23	Anomic	80·6	6	4	2
A-AOS7	63	M	16	Right	Southern	180	Anomic	94·8	9	2	1
A-AOS8	74	F	14	Right	General	13	Anomic	79·5	5	2	2
A-AOS9	50	M	15	Right	Southern	3	Anomic	86·8	6	4	1
A-AOS10	56	F	8	Right	General	44	Conduction	71·9	6	4	2
Mean	56·5		13·2			46·1		74·8	5·2	3·2	1·1
APH1	46	M	8	Right	Southern	8	Wernicke	62·9	8	0	0
APH2	61	M	12	Right	Southern	13	Wernicke	54·0	6	0	0
APH3	59	M	14	Right	Southern	78	Wernicke	68·2	8	0	0
APH4	28	M	10	Left	Southern	91	Wernicke	70·9	6	0	0
APH5	74	F	15	Right	Southern	13	Conduction	64·3	8	0	0
APH6	81	M	12	Right	General	27	Anomic	87·8	9	0	0
APH7	61	M	12	Right	Southern	17	Anomic	96·1	10	0	0
APH8	37	M	12	Right	Southern	8	–	85·8	9	0	0
APH9	67	M	14	Right	Southern	9	–	90·4	9	0	0
APH10	48	M	18	Right	General	22	–	97·8	9	0	0
Mean	56·2		12·7			28·6		77·8	8·2	0	0

Educ. = years of education, MPO = months post-onset, Aphasia = syndrome classification according to the *Western Aphasia Battery* (Kertesz 1982), – indicates that the subject was not classifiable, AQ = Aphasia Quotient from the *Western Aphasia Battery*, Fluency = WAB fluency rating, AOS = apraxia of speech severity rating, Dys. = dysarthria severity rating.

Western Aphasia Battery (WAB) (Kertesz 1982) were given in order to estimate the severity of aphasia and, if possible, classify the aphasia type. The spontaneous speech portion of the testing was audio-recorded and reviewed after the session by the examiner to ensure accurate scoring. As indicated in table 1, the mean aphasia quotients (AQ) were similar across the two patient groups: 74·8 for the subjects with A-AOS and 77·8 for the aphasic subjects. Mean aphasia quotient in the normal group was 99·4 and all normal subjects scored above the 93·8 cut-off point for aphasia on the WAB. One subject in the A-AOS group and two subjects in the aphasia group scored above the cut-off point. All of these subjects considered themselves to have aphasia. They described their difficulties as primarily involving reading and writing and word finding. It was also the clinical judgements of the first and second authors that these patients had aphasia. In the standardization of the WAB, Kertesz (1979) noted that several patients who presented clinically with a mild aphasia did score above the 93·8 AQ cut-off point for normal performance. According to the WAB criteria, five A-AOS subjects were classified as having Broca's aphasia, four were classified as anomic, and one classified as conduction. Four aphasic speakers were classified as having Wernicke's aphasia, one classified as conduction, two classified as anomic, and three were unclassifiable.

A Motor Speech Evaluation (Wertz *et al.* 1984, pp. 99–102) was conducted. This examination was audio-recorded and reviewed by three certified speech–language pathologists, each of whom had over 20 years experience in the differential diagnosis of neurogenic communication disorders. A traditional eight-point rating scale from zero to seven was used to evaluate both dysarthria and apraxia of speech (Collins *et al.* 1983, Hoit-Dalgaard *et al.* 1983). On this scale, zero represented no impairment and seven represented severe impairment. The scale levels between the two end-points were not described further. In order to receive an apraxia of speech diagnosis, the patient was required to present the following speech characteristics (Kent and Rosenbek 1983, Wertz *et al.* 1984):

(1) effortful speech with attempts to self-correct,
(2) frequent articulatory errors, such as substitutions, distortions, omissions, additions, and repetitions,
(3) articulatory variability across repeated productions of the same utterance, and
(4) abnormal prosody.

These criteria were developed and have long been used to identify apraxia of speech in patients who, like our subjects, had a coexisting aphasia. Based on observations of patients with apraxia of speech and minimal language involvement, McNeil and colleagues (McNeil *et al.* 1997) have proposed a different set of criteria that, in several respects, contrast with these classic criteria. For example, it is suggested that patients with apraxia of speech are less variable and more consistent in their sound substitutions than patients with aphasia and no apraxia of speech (e.g. conduction aphasia). However, the applicability of symptoms observed in pure forms of apraxia of speech to the more complex clinical situation of coexisting aphasia and apraxia of speech has not been established and may not be straightforward. Therefore, we elected to use classic, clinical criteria as the basis for our diagnosis.

The three clinicians met for an initial session of observer calibration. During this session, the rating scale was presented and the clinicians discussed which general criteria to use for assignment of different scale levels. Audio-recorded motor speech evaluations of two patients who did not participate in the study (one diagnosed with aphasia and one with aphasia and apraxia of speech) were presented for practice scoring. Following this initial calibration, the three clinicians met during three separate sessions to rate motor speech characteristics of the subjects. Patient and normal subjects were presented in random order during each session and the clinicians were blind to the probable group assignment of each speaker. The clinicians first listened jointly to the audio-recorded motor speech evaluation of each subject. They then independently rated the severity of apraxia of speech and dysarthria, discussed their impressions, and arrived at an agreement about appropriate ratings of both disorders. Inter-observer agreement was calculated on the basis of each clinician's independent rating prior to the group discussion. Initial agreement across all three clinicians in terms of presence or absence of apraxia of speech was 75% for the 20 patient subjects and 100% for the 10 normal subjects. For 100% of the patient subjects, there was agreement among at least two of the clinicians about the presence or absence of apraxia of speech. All three clinicians agreed on the severity rating within one scale level for 60% of the speakers with apraxia of speech and at least two of the clinicians agreed within one scale level for 100% of these speakers. Following group discussion, a consensus about apraxia of speech diagnosis and severity was obtained for all speakers.

All speakers were required to show no or minimal dysarthria. This was operationally defined as a severity rating no higher than two. Although it might have been of theoretical interest to use a stricter criterion and exclude from the sample all subjects with any signs of a potentially coexisting dysarthria, we opted not to do so in order to protect the external validity of the study. In our experience, the majority of patients with A-AOS also have a coexisting mild dysarthria and frequently a right facial hemiparesis. This, of course, is not surprising, given what is known regarding the site of lesion for apraxia of speech. Lesions typically involve posterior–inferior regions of the left frontal lobe in the vicinity of primary motor cortex for the lower face and articulators (Dronkers 1996, Kushner *et al.* 1987, Pellat *et al.* 1991, Schiff *et al.* 1983, Sugishita *et al.* 1987). On the basis of independent ratings, all three clinicians agreed that the criterion of no or minimal dysarthria was met for 100% of the normal subjects and 90% of the patient subjects. According to group consensus, all subjects showed no or minimal dysarthria. As indicated in table 1, no subjects in the aphasia only or the normal groups, but most speakers in the A-AOS group, were judged to have minimal dysarthria, thus confirming our clinical impression of a close association between apraxia of speech and minimal dysarthria.

Normal subjects were recruited by word of mouth. They were friends of the investigators, staff at Vanderbilt University and the Nashville VA Medical Center, and former participants in other research projects at these institutions. All normal subjects showed no evidence and reported no history of neurological pathology. They reported no current or previous speech or language problem. The same tests were administered to the normal subjects as to the patient subjects in order to verify that their speech and language abilities were within normal limits. An effort was made to obtain age, gender, and dialect distributions that approximated those of the patient groups. The mean age for the normal subjects was 62·0 years, which is comparable to 56·5 for the A-AOS group and 56·2 for the aphasia group. The gender distribution was six males to four females for both the A-AOS group and the normal group. However, there was a shortage of referrals of female aphasia subjects. Thus, the gender distribution in the aphasia group was nine males to one female. As the basis for regional dialect classification a three-category system was used (Southern, General American, and New England). Assignment was based on place of residence during childhood and adolescence, patient's own characterization of his/her speech prior to the stroke, and the impressions of the first author. These three criteria were in agreement in all cases but one. One of the normal speakers grew up in the South, but according to herself and the first author used a General American dialect. Therefore, she was assigned to this dialect. The dialect distribution was five Southern to five General American for the normal group and the A-AOS group and eight Southern to two General American for the aphasia group. No subject was classified as using a New England dialect.

Finally, hearing status was evaluated to ensure that the subjects were able to perceive accurately the words presented for repetition in the experimental task. Twenty-seven of the subjects did not wear hearing aids and did not report reduced hearing acuity. They all passed a hearing screening at 40 dB HL in the better ear for the frequencies 1000 and 2000 Hz. This procedure has been recommended for identifying hearing impairment in subjects who are 65 years of age or older (Lichtenstein *et al.* 1988, Ventry and Weinstein 1983). Three subjects (one with A-AOS, A-AOS7, and two with aphasia, APH5 and APH7) regularly used bilateral hearing aids. Their medical records indicated the presence of bilateral high frequency sensorineural hearing loss, characteristic of presbycusis. Two of these subjects wore their aids during the recording. One of the aphasic subjects (APH5) forgot to bring her hearing aids to the recording session and instead used a Williams Sound Pocket Talker personal amplifier that adjusted the speech of the experimenter to a comfortable hearing level. All subjects indicated that they had no trouble hearing what the experimenter said.

Speech samples

Audio-recordings were made in a sound-treated IAC booth. The instrumentation included a high quality reel-to-reel tape recorder (Tascam 22–2) and head-mounted microphone with a constant microphone-to-mouth distance (AKG C410). The productions were digitized from the audiotape and stored on a 486 Gateway 2000 personal computer. The sampling rate was 22 kHz with a low-pass filtering at 10 kHz. Quantization was set at 12 bits.

Single words were printed on white 5″ × 8″ (125 × 205 mm) index cards. An experimenter read these cards and presented them to the subject, one at a time. The subjects were instructed to repeat the words with a normal voice and loudness level. Care was taken to provide sufficient response time. Subjects were allowed to self-correct and these corrections as well as other complete or partial attempts were considered part of each target word. However, there were no prompts to improve accuracy of production. Because subject recruitment and speech sampling took place over the course of several months, it was necessary to use three different experimenters to complete the recordings. All were female graduate students in speech–language pathology, in their 20s, and from the south-eastern region of the United States. They were native speakers of English with no known speech or hearing impairment. Each experimenter elicited productions from both normal and patient subjects. The first author was present during all recordings to operate the recording equipment and to ensure that similar instructions and procedures were used with all subjects.

The subjects were asked to produce 100 single words from a published intelligibility test (Kent *et al.* 1989). All words were monosyllabic and often minimally different from each other. Each subject produced the same set of 70 target words from the multiple choice version of the test. In addition, 30 phonetically similar words were recorded prior to the intelligibility words. The phonetically similar words were randomly selected for each subject from the foils used in the test. Thus, these words formed a unique set for each speaker and served as dummy words to reduce learning effects during the perceptual testing. The 30 dummy words were produced first in random order. Next, the 70 intelligibility words were recorded. These words were randomized once and produced in the same random order by all subjects. A second production of the intelligibility words was obtained from all speakers in order to determine test–retest reliability. This occurred on the same day as the first production, following a lunch break and/or a session of speech recordings that were not part of the present investigation. During this second recording, the 70 intelligibility words were elicited again, along with 30 new dummy words.

Intelligibility testing

The words were presented to 30 normal listeners for identification. The mean age of these listeners was 31 years and there was an approximately equal distribution across male and female subjects. Half of the subjects had some prior phonetic training as students of speech and hearing science. Of these, five were speech–language pathologists and therefore had some experience in neurogenic communication disorders. Phonetically trained listeners and speech–language pathologists were distributed equally across listening conditions. All listeners demonstrated normal hearing on the basis of a pure-tone screening at 20 dB HL for the octave frequencies between 500 and 4000 Hz. The testing was conducted in a sound-treated IAC booth with the words presented over headphones at a comfortable listening level of approximately 80 dB SPL. One speaker at a time was presented, with the 70 intelligibility words and 30 dummy words mixed and randomized separately for each speaker. Self-corrections and repeated partial and complete attempts at the target word were included in the presented utterances, as long as they were not separated by more than one second. The listeners were told that they would hear single words produced by several speakers and that some of the speakers had suffered a stroke. They were instructed to write down each word they thought the person was trying to say and to guess if they were not sure. Listeners controlled the onset of each stimulus by pressing a button but were only able to listen to each stimulus once. The average testing time for each listener was about 90 minutes. However, because each listener was able to control the rate of presentation, there was some variation among individual listeners. In order to minimize fatigue, the listeners were required to take at least one break during the testing and encouraged to take additional breaks whenever they wanted.

The listeners were divided into three groups with 10 listeners in each group. Each listener transcribed productions from 10 speakers distributed across the three speaker groups. Additionally, productions from two speakers were randomly selected for each listener group to be presented a second time in order to estimate intra-observer reliability. Retest productions from two other speakers were also included in the transcription process in order to estimate test–retest reliability. Thus, across all listeners combined, there were 12 reliability repetitions: six to measure intra-observer reliability and six to measure test–retest reliability. The intra-observer and test–retest reliability presentations were mixed and randomized with the experimental presentations, with the only exception that at least two speakers were presented between the repeated presentations.

Including all experimental and reliability presentations, each listener transcribed a total of 1400 words. Because the same set of words was used for all speakers, five strategies were adopted to minimize learning effects and their potential influence on the intelligibility scores:

(1) Each listener transcribed the sample only 14 times (experimental productions from 10 speakers and reliability productions from four speakers).
(2) The order of the words was randomized separately for each speaker.
(3) In addition to the 70 words for the intelligibility test, 30 dummy words that were phonetically similar to several of the target words and unique for each speaker were presented.
(4) Listeners were informed that many of the words were minimally different from each other in terms of their phonemic content.
(5) For each listening group, the speakers were presented in five different random orders.

An overall intelligibility score was calculated as the number of correctly transcribed words divided by the total number of words. The dummy words were disregarded in this calculation. Homophones, such as 'new' and 'knew' and 'write' and 'right' were considered interchangeable, and minor spelling errors that did not change the phonemic form of the utterance were scored as correct. The mean score for a given speaker across the 10 listeners served as the dependent variable for intelligibility.

Results

Reliability

Three types of reliability were estimated: inter-observer, intra-observer, and test–retest reliability. Inter-observer agreement is reported in table 2. As the first measure, average intra-speaker standard deviation was obtained. This was done by calculating the standard deviation across listeners for each speaker (see table 4) and then computing the mean across all speakers in the group. The average standard deviation ranged from 2·90 in the normal group to 4·62 in the A-AOS group. As an additional measure of inter-observer reliability, the percentage of individual observations that were within 5 and 10 percentage points of the listener group mean was calculated. As is shown in table 2, at least 75% of all observations were within 5 percentage points of the listener group mean for each of the speaker groups.

As a measure of intra-observer reliability, the Pearson product moment correlation was calculated between the experimental and the reliability observations. This correlation was 0·95. The mean score for the experimental observations was 84% and for the repeated observations 85%. As expected, a

Table 2. Inter-observer reliability for the three speaker groups and for all speakers combined

Group	Intelligibility	Average SD	Within 10%	Within 5%
Normal	95·2	2·90	99	91
A-AOS	73·8	4·62	97	75
Aphasia	85·4	3·38	97	91
All	84·8	3·63	98	86

Table entries indicate mean intelligibility score, average intra-speaker standard deviation, and percentage of individual observations that were within 5 and 10 percentage points of the listener mean.

Table 3. Intra-observer and test–retest reliability for the three speaker groups and for all groups combined

	Intra-observer reliability ($r = 0·95$)		Test–retest reliability ($r = 0·98$)	
Group	Within 10%	Within 5%	Within 10%	Within 5%
Normal	100	95	100	95
A-AOS	90	70	100	55
Aphasia	100	100	100	90
Mean	97	88	100	80

Table entries indicate percentage of observations for which the overall scores of the experimental and repeated observations corresponded within 5 and 10 percentage points. The Pearson product moment correlation between experimental and repeated presentations are shown in parenthesis for each type of reliability.

paired *t*-test showed that this difference was not significant. The percentage of observations for which the overall scores of the experimental and repeated presentations corresponded within 5 and 10 percentage points was calculated. As indicated in table 3, the original and repeated scores were within 5 percentage points of each other at least 70% of the time for each of the speaker groups. Moreover, the mean intelligibility scores across all 10 listeners were within two percentage points for five of the six speakers used to calculate intra-observer reliability and within 5 percentage points for all six of these speakers.

Test–retest reliability was also estimated by computing the Pearson product moment correlation between the scores of the test and retest presentations. This correlation was 0·98. The mean score for the primary observation was 84·2% and for the repeated observation 84·0%. A paired *t*-test showed that this difference was not significant. All test–retest scores were within 10 percentage points of each other across the speaker groups, and scores were within 5 percentage points for at least 90% of the normal and aphasia only speakers (see table 3). However, only 55% of the observations were within 5 percentage points of each other for the A-AOS speakers. The mean scores across the 10 listeners were within two percentage points for all of the six test–retest presentations.

Group differences

The performance of the three speaker groups is shown in figure 1. The percentage of intelligible words was lowest ($M = 73·8$) and most variable ($SD = 16·5$) in the

Table 4. Intelligibility scores for individual speakers

Speaker	Intelligibility	Speaker	Intelligibility	Speaker	Intelligibility
NOR1	97·5 (1·96)	A-AOS1	39·4 (6·87)	APH1	71·0 (7·26)
NOR2	97·0 (2·16)	A-AOS2	85·1 (4·09)	APH2	88·6 (2·55)
NOR3	96·5 (2·63)	A-AOS3	77·4 (3·63)	APH3	89·3 (2·87)
NOR4	97·8 (2·78)	A-AOS4	90·2 (2·70)	APH4	88·2 (3·76)
NOR5	93·9 (3·28)	A-AOS5	92·4 (2·76)	APH5	61·0 (3·97)
NOR6	88·5 (6·31)	A-AOS6	70·9 (7·84)	APH6	88·1 (3·72)
NOR7	96·0 (1·76)	A-AOS7	66·4 (4·14)	APH7	93·6 (2·37)
NOR8	94·7 (2·58)	A-AOS8	89·4 (4·14)	APH8	88·8 (2·78)
NOR9	97·7 (2·71)	A-AOS9	62·7 (5·44)	APH9	91·3 (2·06)
NOR10	92·3 (2·79)	A-AOS10	63·9 (4·58)	APH10	93·8 (2·48)

NOR = normal, A-AOS = aphasia and apraxia of speech, APH = aphasia only. Standard deviations across the 10 listeners are listed in parentheses.

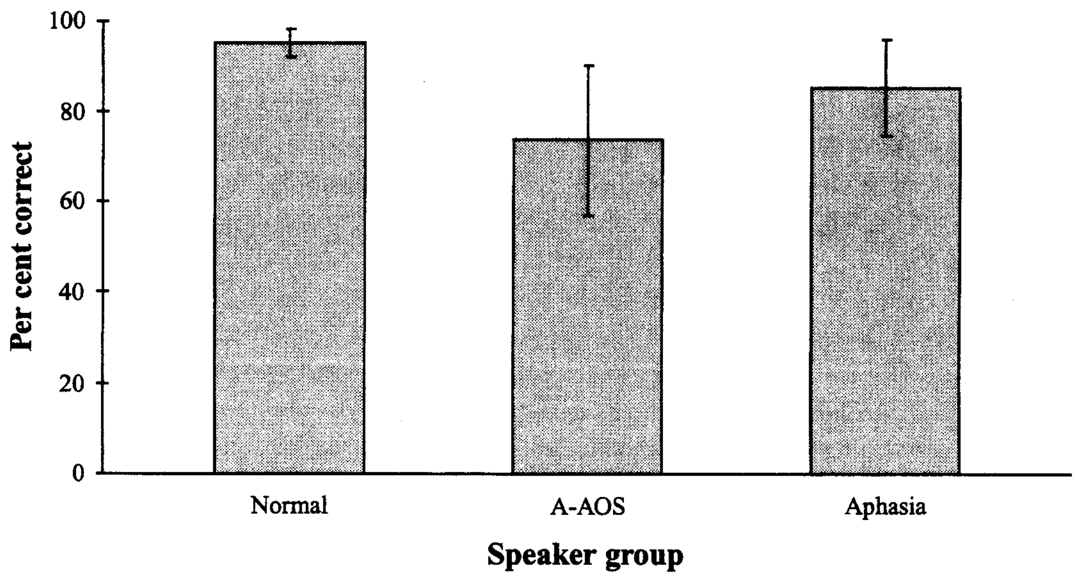

Figure 1. Single word intelligibility (per cent) for the three speaker groups. Bars represent standard deviations.

A-AOS group, intermediate in the aphasia group ($M = 85{\cdot}4$, $SD = 10{\cdot}7$), and highest ($M = 95{\cdot}2$) and least variable ($SD = 3{\cdot}0$) in the normal group. The Kruskal–Wallis test revealed a significant overall difference [$\chi T^2(2) = 16{\cdot}24$, $p < 0{\cdot}0005$] among the group means. Follow-up analyses showed significant differences between the normal group and the aphasia group ($U = 8{\cdot}0$, $p < 0{\cdot}005$) and between the normal group and the A-AOS group ($U = 4{\cdot}0$, $p < 0{\cdot}001$), whereas the difference between the aphasia and A-AOS groups was not significant.

In order to evaluate the validity of impaired speech intelligibility as a feature of A-AOS and aphasia, the scores for individual speakers were examined. These values are listed in table 4 and plotted in figure 2 for further illustration. As can be seen, there is relatively limited overlap between the intelligibility scores of the normal and the A-AOS groups. Seven of the 10 speakers with A-AOS obtained intelligibility scores below the lowest normal score. There was no straightforward

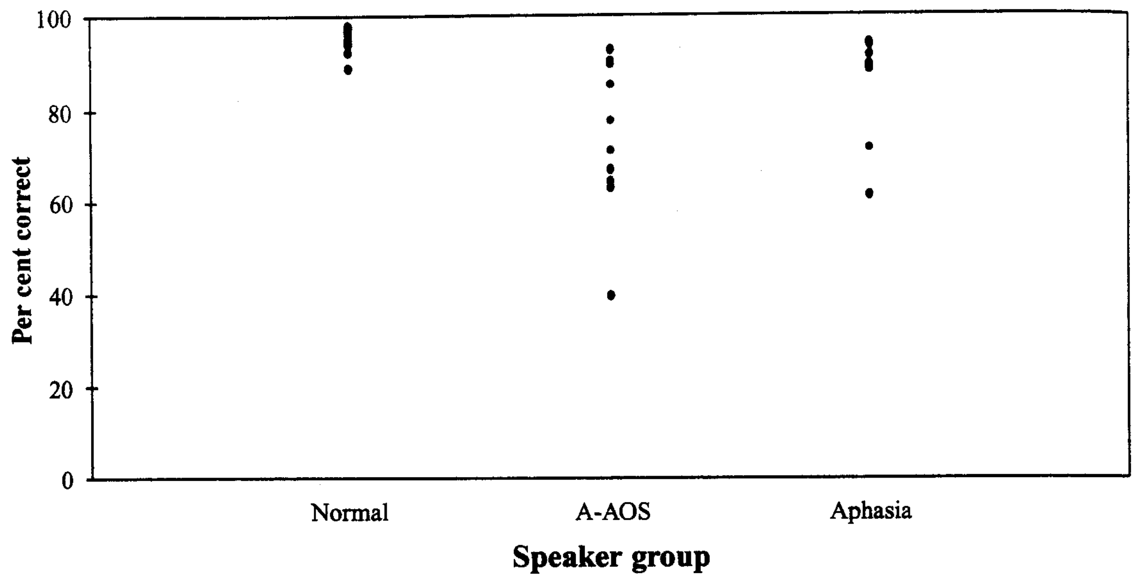

Figure 2. Single word intelligibility (per cent) for individual speakers across the three groups.

relationship between intelligibility and aphasia type or severity. The Pearson product moment correlation between intelligibility scores and the Aphasia Quotient on the WAB was close to zero ($r = 0{\cdot}06$) for these subjects with A-AOS. The most as well as the least intelligible subjects included both patients with Broca's aphasia and patients with anomic aphasia. There also did not seem to be a clear correlation between presence/absence of minimal dysarthria and intelligibility. Although the limited sample size precludes any definite conclusions, the observed patterns indicate that A-AOS typically is associated with reduced speech intelligibility, at least as measured on a test of single word production.

In contrast to the relatively clear difference between the normal and A-AOS speakers, there was substantial overlap in intelligibility scores between the aphasia group and both the normal and the A-AOS groups. Eight of the 10 aphasic speakers obtained intelligibility scores within the range of the normal speakers and the remaining two (APH1 and APH5) obtained scores that were in the mid to lower range of performance of the A-AOS subjects. This heterogeneity reflects what we expect to find in the general aphasic patient population. Different patients show different patterns of impairment; production difficulties affect accuracy of articulation more in some patients than in others (Blumstein 1973, Halpern *et al.* 1976). The two patients that scored the lowest on the intelligibility test also obtained Aphasia Quotient scores that were in the lower range for the group. One was diagnosed with Wernicke's aphasia and the other with conduction aphasia. Subjectively, these two patients produced a greater number of speech sound substitutions during the speech and language evaluation than most other subjects in the aphasia group. The Pearson product moment correlation between intelligibility scores and Aphasia Quotient for the aphasic group was moderate ($r = 0{\cdot}57$) and approached ($p < 0{\cdot}10$), but did not reach, statistical significance. Taken together, the results thus indicate that some patients who have aphasia without A-AOS present with reduced speech intelligibility of single word production, whereas many others produce speech of normal intelligibility.

In order to examine the degree of association between the intelligibility testing and perceived overall severity of A-AOS, the Pearson product moment correlation was computed between intelligibility scores and A-AOS severity ratings in the A-AOS group. The correlation was moderate ($r = -0.58$) and approached, but did not reach, statistical significance ($p < 0.10$). Inspection of the data for individual subjects supports this limited correlation. Although the subject that was rated as most severe obtained the lowest intelligibility score and the subject rated as the mildest received the highest intelligibility score, there was also quite a bit of variability within given severity levels. For example, among the five speakers that were assigned a severity score of 4, intelligibility ranged from 63% to 90%.

Discussion

The data provide answers to each of our questions. First, at least on the basis of the relatively limited number of reliability repetitions we were able to present in the present investigation, it appears that intelligibility testing can be conducted reliably for speakers with A-AOS and speakers with aphasia. However, whereas judge reliability was acceptable for both groups, test–retest reliability was more questionable for A-AOS speakers. We found that test–retest reliability was poor for individual listeners' transcriptions of speech produced by speakers with A-AOS. Again, the number of subjects on which the reliability data are based was limited and probably should be interpreted with some caution. It was reassuring that test–retest reliability for the listening group as a whole was quite good. It will be important in future research to verify the observed reliability levels across a greater number of speakers and listeners. Particularly needed are complete reliability data that include several speakers with A-AOS and aphasia. Reliability data for listening groups as a whole are also needed, since intelligibility scores typically reflect averages across such a group, rather than scores from individual listeners. Finally, more detailed test–retest information is particularly needed for speakers with moderate to low intelligibility, where inconsistency may be a special problem.

Second, both A-AOS and aphasia are associated with reduced levels of single word intelligibility. This type of intelligibility assessment seems to be particularly sensitive to speech deviations in the A-AOS population, where the majority of speakers obtained scores well below the normal range. On the average, there was no difference between the level of intelligibility in A-AOS and the level of intelligibility in aphasia. However, it should be noted that our data showed that speech intelligibility among patients with aphasia and no coexisting apraxia of speech is quite variable, with some patients performing within the normal range and others having substantially reduced intelligibility. Subjectively, it was the patients that presented with the greatest frequency of phonemic paraphasic errors that were the least intelligible. Although not surprising, this finding indicates that it may be important to pay attention to differences in intelligibility when comparing phonetic deviation across subjects within as well as across traditional aphasia syndromes. For example, a finding of more deviant acoustic properties in patients with A-AOS than in patients with aphasia only (Baum *et al.* 1990, Blumstein *et al.* 1980, Itoh *et al.* 1982) may at least in part be explained on the basis of greater overall speech involvement among speakers with apraxia of speech than among patients who present with aphasia only. Similarly, different levels of intelligibility may also be associated with individual phonetic differences within speaker groups.

Third, there is no substantive relationship between single word intelligibility and severity of aphasia in speakers with A-AOS and a moderate, but non-significant, relationship between single word intelligibility and aphasia severity in patients without apraxia of speech. This confirms our impression that speech production errors in speakers with A-AOS are determined primarily by a disorder separate from, but coexisting with, an aphasia of varying type and severity. In contrast, articulatory errors produced by individuals who have aphasia and no apraxia of speech are more closely related to aphasia severity.

Fourth, there is a moderate, but non-significant, negative relationship between monosyllabic single word intelligibility and clinical ratings of apraxia of speech severity. It appears that at least some of what goes into a judgement of apraxia of speech severity is based on this aspect of speech intelligibility. Therefore, in A-AOS, as in other speech disorders, single word intelligibility may be a viable way to quantify severity of involvement. However, it should be noted that the correlation was only moderate and not statistically significant. Besides the small sample size, there are a number of potential reasons for this moderate correlation. For example, the severity ratings may have been of limited accuracy and precision, due to the subjective nature of the rating scale. Furthermore, the spread of severity levels was limited. Half of the subjects obtained a rating in the middle of the scale, at level 4. Finally, the degree of association should not be surprising given the many differences in speech sample and listening task for the two measures. Of the different characteristics of A-AOS, one may hypothesize that articulatory errors, such as substitutions, distortions, additions, omissions, and repetitions were most relevant to the intelligibility test. Single word intelligibility is probably more likely to be influenced by these properties than by any other clinical features of the disorder. In contrast, deviant prosody, inconsistency across repeated productions, and articulatory breakdown with longer utterances would not be apparent in single word productions. Thus, clinical severity ratings and single word intelligibility scores are likely to highlight slightly different properties of A-AOS. Therefore, whereas both provide a general measure of speech involvement, they are suitable for different purposes. For example, single word intelligibility testing may have greater validity for analysis of segmental aspects of speech production, whereas overall clinical ratings may be more sensitive to suprasegmental factors.

Although preliminary, the results are promising. As noted, the possibility of obtaining reliable speech intelligibility measures for speakers with A-AOS has obvious clinical implications. Direct assessment of speech intelligibility offers a different perspective and several advantages to traditional articulatory testing. The inherently functional definition of these measures, objective scoring procedures, and potential ability to serve as a measure of overall speech involvement are among the most important ones. Ideally, an intelligibility test can also provide information about the reasons why a person obtains a particular score and which phonetic variations are most likely to reduce a person's communicative effectiveness. Thus, patient management may not only be monitored, but also guided by an intelligibility test. The Kent *et al.* (1989) protocol was developed with this particular objective in mind. The target words differ from each other across a number of carefully varied, quantifiable phonetic contrasts. We are currently analysing the phonetic error patterns that the subjects in this study made. The results of this analysis may contribute to a better understanding of which phonetic deviations in A-AOS are functionally important to listeners. There are many ways in which perceptual, acoustic, and physiologic analyses of the speech signal can and should complement each other. Single word speech intelligibility can serve an important function by helping to bridge the gap between acoustic and perceptual characterizations of apraxia of speech.

References

Baum, S. R., Blumstein, S. E., Naeser, M. A. and Palumbo, C. L. 1990, Temporal dimensions of consonant and vowel production: an acoustic and CT scan analysis of aphasic speech. *Brain and Language*, **39**, 33–56.

Blumstein, S. E. 1973, *A Phonological Investigation of Aphasic Speech* (The Hague: Mouton).

Blumstein, S. E. Cooper, W. E., Goodglass, H., Statlender, S. and Gottlieb, J. 1980, Production deficits in aphasia: a voice onset time analysis. *Brain and Language*, **31**, 153–170.

Collins, M., Rosenbek, J. C. and Wertz, R. T. 1983, Spectrographic analysis of vowel and word duration in apraxia of speech. *Journal of Speech and Hearing Research*, **26**, 224–230.

Culton, G. L. 1969, Spontaneous recovery from aphasia. *Journal of Speech and Hearing Research*, **12**, 825–832.

Dronkers, N. F. 1996, A new brain region for coordinating speech articulation. *Nature*, **384**, 159–161.

Gandour, J., Dechonkit, S., Ponglorpisit, S., Khunadorn, F. and Boongird, P. 1993, Intraword timing relations in Thai after unilateral brain damage. *Brain and Language*, **45**, 160–197.

Halpern, H., Keith, R. L. and Darley, F. L. 1976, Phonemic behavior of aphasic subjects without dysarthria or apraxia of speech. *Cortex*, **12**, 365–372.

Hoit-Dalgaard, J., Murry, T. and Kopp, H. G. 1983, Voice onset time production and perception in apraxic subjects. *Brain and Language*, **20**, 329–339.

Itoh, M., Sasanuma, S., Tatsumi, I. F., Murakami, S., Fukasako, Y. and Suzuki, T. 1982, Voice onset time characteristics in apraxia of speech. *Brain and Language*, **17**, 193–210.

Johns, D. F. and Darley, F. L. 1970, Phonemic variability in apraxia of speech. *Journal of Speech and Hearing Research*, **13**, 556–583.

Kent, R. D. and Rosenbek, J. C. 1983, Acoustic patterns of apraxia of speech. *Journal of Speech and Hearing Research*, **26**, 231–249.

Kent, R. D., Miolo, G. and Bloedel, S. 1994, The intelligibility of children's speech: a review of evaluation procedures. *American Journal of Speech–Language Pathology*, **3**, 81–93.

Kent, R. D., Weismer, G., Kent, J. F. and Rosenbek, J. C. 1989, Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, **54**, 482–499.

Kertesz, A. 1979, *Aphasia and Associated Disorders: Taxonomy, Localization, and Recovery* (New York: Grune & Stratton).

Kertesz, A. 1982, *Western Aphasia Battery* (New York: Grune & Stratton).

Kushner, M. Reivich, M., Alavi, A., Greenberg, J., Stern, M. and Dann, R. 1987, Regional cerebral glucose metabolism in aphemia: a case report. *Brain and Language*, **31**, 201–214.

Lichtenstein, M. J., Bess, F. H. and Logan, S. A. 1988, Validation of screening tools for identifying hearing impaired elderly in primary care. *Journal of the American Medical Association*, **259**, 2875–2878.

McNeil, M. R., Robin, D. A. and Schmidt, R. A. 1997, Apraxia of speech: definition, differentiation, and treatment. In M. R. McNeil (Ed.) *Clinical Management of Sensorimotor Speech Disorders* (New York: Thieme Medical), pp. 311–344.

Odell, K., McNeil, M. R., Rosenbek, J. C. and Hunter, L. 1990, Perceptual characteristics of consonant production by apraxic speakers. *Journal of Speech and Hearing Disorders*, **55**, 345–359.

Osberger, M. J. 1992, Speech intelligibility in the hearing impaired: research and clinical implications. In R. D. Kent (Ed.) *Speech Intelligibility in Speech Disorders: Theory, Measurement, and Management* (Amsterdam: Benjamins), pp. 233–264.

Pellat, J., Gentil, M., Lyard, G., Vila, A., Tarel, V., Moreau, O. and Benabid, A. L. 1991, Aphemia after a penetrating brain wound: a case study. *Brain and Language*, **40**, 459–470.

Schiff, H. B., Alexander, M. P., Naeser, M. A. and Galaburda, A. M. 1983, Aphemia: clinical–anatomic correlations. *Archives of Neurology*, **40**, 720–727.

Seddoh, S. A., Robin, D. A., Sim, H., Hageman, C., Moon, J. B. and Folkins, J. W. 1996, Speech timing in apraxia of speech versus conduction aphasia. *Journal of Speech and Hearing Research*, **39**, 590–603.

Square, P. A., Darley, F. L. and Sommers, R. K. 1982, An analysis of the productive errors made by pure apractic speakers with differing loci lesions. In R. H. Brookshire (Ed.) *Clinical Aphasiology: Conference Proceedings* (Minneapolis, MN: BRK), pp. 245–250.

Sugishita, M., Konno, K., Kabe, S., Yunoki, K., Togashi, O. and Kawamura, M. 1987, Electropalatographic analysis of apraxia of speech in a left hander and in a right hander. *Brain*, **110**, 1393–1417.

Ventry, I. M. and Weinstein, B. 1983, Identification of elderly people with hearing problems. *ASHA*, **25**, 37–42.

Wertz, R. T., Lapointe, L. L. and Rosenbek, J. C. 1984, *Apraxia of Speech in Adults* (San Diego, CA: Allyn & Bacon).

Yorkston, K. M., Dowden, P. A. and Beukelman, D. T. 1992, Intelligibility as a tool in the clinical management of dysarthric speakers. In R. D. Kent (Ed.) *Intelligibility in Speech Disorders: Theory, Measurement, and Management* (Amsterdam: Benjamins), pp. 265–286.