Review

Issues in perceptual speech analysis in cleft palate and related disorders: a review

Debbie Sell
Department of Speech & Language Therapy, Great Ormond Street Hospital for Children, London, UK

(Received 27 April 2004; accepted 1 September 2004)

Abstract

Background: Perceptual speech assessment is central to the evaluation of speech outcomes associated with cleft palate and velopharyngeal dysfunction. However, the complexity of this process is perhaps sometimes underestimated.

Aims: To draw together the many different strands in the complex process of perceptual speech assessment and analysis, and make recommendations for practice.

Main Contribution: This review examines issues such as data sampling, data collection/recording, archiving, the advantages and disadvantages of lay and specialist listeners, approaches to data analysis, reliability, and data interpretation.

Conclusions: The need to capture meaningfully the sound of speech based on detailed phonetic analysis is argued. Also described are some of the differences in measuring speech for clinical, audit and research activities. Blind independent analysis of speech data by specialist therapists is recommended as the gold standard methodological approach when reporting audit and research outcomes. The requirement for ongoing training in listening skills for specialist therapists is advocated. The limitations of an impairment-based-only approach to measurement are also illustrated, indicating the need to develop outcome measures that incorporate more functional issues that affect quality of life.

Keywords: perceptual speech assessment, data, audio and video recording, reliability, frameworks of speech analysis, clinical audit, data interpretation, randomized blind independent analysis or assessment.

Background

Speech is considered one of the primary outcome measures of cleft lip and palate (CLP) management, and yet its measurement is elusive and challenging. It has been
well recognized for many years that there is a need for one international approach to perceptual speech analysis (McComb 1989, Henningsson and Hutters 1997, Sell 2001, Kuehn et al. 2002). Indeed, one of the aims of the 1993 International Congress on Cleft Palate and Related Craniofacial Anomalies was to establish basic international protocols of speech assessment and measurement. Although different approaches have been proposed, there still remains no consensus (Henningsson and Hutters 1997, Hirschberg and Van Demark 1997), but it is encouraging that Kuehn et al. (2002) have reported on work in progress to address this issue.

Despite this challenge, perceptual speech assessment is recognized to be of the utmost importance. Gerratt et al. (1991) described it as being at the core of the speech and language therapy profession and the standard against which instrumental measures are validated (McWilliams et al. 1984, Dalston and Warren 1986, Haapanen 1991, Hirschberg and Van Demark 1997, Sweeney and Sell, submitted). McWilliams et al. (1990) called for reliable and detailed perceptual speech data substantiated by instrumentation. Indeed, Kuehn and Moller (2000) concluded that there was no instrumental technique that could replace the use of perceptual judgements, in part because the measurement technique needs to be practical and appropriate for use with young children. Howard and Heselwood (2002) also argued that perceptual analysis is indispensable, whilst recognizing the importance of instrumental analysis. Kuehn and Moller (2000) stated that more standardized descriptions of speech parameters, methods and procedures that were repeatable, with demonstrated reliability of listener judgements were required. Gooch et al. (2001) also emphasized the need for researchers to include specific details of listener training methods, conditions of transcription tasks, procedures for calculating reliability, and also emphasized standardization of procedures. Keuning et al. (1999) urged for further investigation of the factors that may affect perceptual assessment. In the area of voice, Kreiman et al. (1993) have drawn attention to the influence of the type of rating scales on reliability and statistical analysis of data, and yet in the field of CLP, relatively little has been written about the process, except for the welcomed recent contribution by Lohmander and Olsson (2004). Indeed, almost all the focus in the CLP literature has been on the framework of speech analysis, and yet there are many other aspects, some of which have been listed above, that need to be considered in order to have datasets that can be meaningfully compared. Lohmander and Olsson, in their critical review of methodology in the perceptual assessment of speech in patients with cleft palate, reported concerns regarding method of data collection, type and documentation of the data, reliability and method of measurement.

Aims
Grunwell et al. (1993) described the general principles of perceptual speech assessment in the clinical context as data, recording, analysis and interpretation. The purpose of this paper is to review these areas, and to draw together where possible recommendations for practice. This paper makes explicit many of the issues that are relevant to clinicians concerned with measuring speech outcomes, particularly given, the UK Government requirement for inter-centre audit to monitor outcomes regularly and to thereby ensure standards of care in CLP management (CSAG 1998).
Before doing so, however, it is appropriate perhaps to consider at this point the term ‘clinical context’. As clinicians, there are multiple functions that are subsumed under a clinical function, including assessment, screening, treatment planning, delivery, monitoring, and evaluation. More recently, clinical audit or quality improvement has become an integral part of clinical practice too. This is defined by NICE (2002) as ‘a quality improvement process that seeks to improve patient care outcomes through systematic review of care against explicit criteria and the implementation of change’. Health professionals are encouraged to look at what they are doing against agreed standards and make changes to practice (DoH 2000). Generally, compared with research studies, clinical audit studies can be less rigorous in their design with more tolerance of bias and heterogeneity of subjects, but it should be appreciated that results are less generalizable. As in research, there is still the need for a clearly defined question or objective, with explicit, reliable and valid measurable outcomes. There are, therefore, outcomes associated with clinical functions, clinical audit and research. A word of caution is needed regarding the use of live clinical data as audit or research outcome data. Increasingly with the proliferation and sophistication of database technology, outcome data are entered into a clinical database, and yet it should be appreciated that such data are susceptible to bias. Mehendale and Sommerlad (2003) describe how the results of speech assessments recorded in the health record at the time of a clinical review may be influenced by positive or negative comments of the patient or the parents. Knowledge of the rater as to the nature of the datum point (e.g. pre- or post-secondary surgery) may also be a source of bias. Clinical databases are excellent for recording the characteristics of a population and processes of care, but it is important to be aware of the standing of such clinical outcome data. This does not mean that these types of data should never be collected and reported. It can be usefully used to inform clinical practice within a unit (Pereira et al. 2002, Van der Ham et al. 2002). However, the results should be interpreted as the local evidence of a unit and should not be considered generalizable.

**Main contribution**

Each of the areas of data, recording, analysis and interpretation are now reviewed related to the cleft palate literature but drawing on other areas of speech pathology where transferable knowledge may be applicable. Issues around the reliability of speech assessment in cleft palate are discussed given its central importance. There has also been some advocacy for professionals other than speech and language therapists to evaluate speech outcome in individuals with cleft palate, and this literature will be reviewed. Finally, the focus on an impairment-based approach to analysis is also considered.

**Data**

It is usual for speech sampling in this field to include stimulability, rote speech, sentence and syllable repetition, and a sample of conversational speech. Grunwell et al. (1993) stated that a spontaneous or conversational speech sample is important. Kuehn and Moller (2000) detailed that the conversational speech sample may provide important information about consistency or deterioration of articulation.
proficiency and changes in resonance characteristics. Nevertheless, sentence sampling is not only an expedient technique in this population but also allows for control of the phonetic content of the elicited speech sample. Van Demark (1964) reported a high correlation between a task of sentence repetition and spontaneous speech. Importantly it allows for comparable speech samples to be routinely collected as part of data collection on a longitudinal basis. Each language should have a nationally agreed speech sample including a set of sentences. Sell et al. (1999) proposed that each sentence should have one target consonant only in different word positions. Each sentence should be formulated based on phonetic principles with no other influencing ‘vulnerable’ consonants in the sentence, and controlled for the possible effects of assimilation. Furthermore, it should be possible for the sentences to be represented pictorially, be meaningful and relevant, containing the maximal number of each target consonant. Lohmander and Olsson’s review found that evaluation was most frequently based on spontaneous speech and single words, and only 9% of articles were based on sentence repetition. Lohmander and Olsson (2004) also highlighted the growing interest in cross-linguistic studies, and yet such studies produce their own challenges. Hutters and Heningsson (2001, 2005) have drawn attention to the importance of having phonetically similar speech material, for example consonant–vowel syllables and carefully constructed sentences, in order to ensure there are comparable datasets. The Eurocleft Speech Group (2000) advocated that a cross-linguistic comparison should be based on the common consonants in the languages. These recommendations clearly make narrow phonetic transcription mandatory.

There is some controversy over the method of sentence elicitation. Van Demark (1964) found that consonant production was better for repeated sentences than read sentences. More recently, Keuning et al. (1999) reported no differences between read and repeated sentences. There are some advantages, however, to using repetition of sentences. Primarily the task does not interact with literacy skills, and sentence repetition can be used as young as four years, sometimes even younger, in contrast to reading. In addition, analysis, be it live for clinical purposes, or subsequently undertaken from recordings, is made easier, as the rate at which sentence elicitation takes place is under the control of the adult eliciting the sentences (John et al. 2003). Furthermore, unlike in a reading task, the patient can be encouraged to maintain eye contact with the listener in order that the face can be viewed during analysis. It is important that one approach is adopted and maintained longitudinally over time in order to keep data sets comparable, even when literacy skills have been acquired. It is always possible to add additional speech tasks as long as the core dataset is always gathered. Kuehn and Moller (2000), for example, advocated the use of standard reading passages for older patients.

Recording

There has been a long tradition of audiotape recordings in the clinical context for purposes of documentation. With the advent of video recordings, the use of this medium has been adopted. A review of the Cleft Palate–Craniofacial Journal between 1997 and 2002 revealed eight speech articles, five of which were analysed from audio speech samples and three from video samples, and yet there are conflicting findings in the literature to inform how these two mediums impact analysis. McGurk and
MacDonald (1976) provided evidence that speech perception of normal speech is influenced by visual clues. Podol and Salvia (1976) also showed that ‘seeing’ patients with CLP influenced ratings, but in contrast, Ramig (1982) and Moller and Starr (1984) found there was no difference. Although Ramig (1982) emphasized the need for future research, surprisingly no further studies have been published. There has been a recent trend in the UK to base judgements on video recordings (Sell and Grunwell 1990, Eurocleft Speech Group 2000, Sell et al. 2001). Recently Sell et al. (2002) reported on an unpublished study to investigate differences in judgements between the audio and video analyses of the speech of children with CLP, using the national Clinical Standards Advisory Group Study audio and video speech recordings archive. They found there were no statistically significantly differences between the two media, although there was a trend for video analysis to produce more critical ratings of speech consonant characteristics, hypernasality and nasal turbulence. Although this suggests that it is possible to compare outcomes based on these two different media, there were methodological flaws in the study. The video recordings were recorded on analogue and not digital tape, in contrast to the audio recordings. A reliability study of only two trained listeners was undertaken based on the digital audio recordings alone. It was not certain if the whole range of speech characteristics, and in particular their co-occurrence, were represented. It is, therefore, important for further studies to be undertaken in this area, and as yet, the influence of the medium of data recording is unknown. Lohmander and Olsson (2004) consider that the choice of medium depends on the purpose of the study but do not expand on this.

Before leaving this issue, however, the importance of the recordings archive needs to be stressed. It is only through such data that independent assessment by specialist speech and language therapists based on randomized order of presentation of recordings is possible (Wyatt et al. 1996). It is recommended that this should be the standard used in audit and research projects. It is of concern that Lohmander and Olsson (2004) reported a large number of studies based on live recordings, with the drawback of inherent bias, as previously discussed. The use of audio and video recordings brings further challenges. Adequate systems for documentation of recordings are of paramount importance in order that retrieval is easily undertaken. For example, in studies of surgical outcome for speech, longitudinal pre- and post-operative tape recordings for individual patients need to be easily accessed in order for them to be edited for subsequent independent listening (Sommerlad et al. 2002). Gooch et al. (2001) also discussed the importance of the quality of the recording and listening environments, and the need to ensure uniformity of the amplitude of speech samples. John et al. (2003) also suggested that the nature of the speech-recording medium, such as analogue or digital, may have an impact on analysis and this requires investigation.

Speech analysis

In turning attention to agreeing a framework of speech analysis there are many challenges, and many different approaches have been proposed. This is not just a feature of this field. Kreiman et al. (1993) identified 57 different perceptual voice schemes in use in the USA! The approach adopted depends on several factors: whether measurement is taking place for clinical, audit, or research purposes, the
population under study, the question being asked, and the types of outcomes required. It will be recognized that these are the principles of evidence-based practice (Sackett et al. 1998). For example, is a detailed phonetic description of consonant production required, or an approach to capture the developing sound system of infants or young toddlers, or perhaps is an overall judgement of speech quality/intelligibility, or improvement or elimination of features in an older group an appropriate approach to adopt? In the Eurocleft population there was a need for a detailed quantitative and qualitative framework of phonetically defined errors, whereas an investigation into the impact of late palate repair in adolescents and adults with unrepaired CLP required a very different approach but one which was still based on phonetic principles (Eurocleft Speech Group 1994, 2000, Sell and Grunwell 1990). Nevertheless, for similar types of populations and questions, e.g. measuring speech performance at 5 years of age for audit purposes, or reporting outcome of secondary surgery for speech, one approach should ideally be adopted.

In the main, early speech reporting systems were simplistic and inadequate and developed by professionals other than speech and language therapists (Sell and Grunwell 2001). A system that still is in usage is the traditional error framework of substitution, distortion, omission, and gross substitution, which has been severely criticized by linguists (Grunwell et al. 1993), and ignores the major impact phonology has had on the understanding of speech disorders generally (Russell and Harding 2001).

More recent approaches to measurement have included the Pittsburgh Scale (McWilliams and Phillips 1979), the Categorical System of Articulation Problems in Cleft Palate (Ainoda and Okazaki 1993) and the Temple Street Scale (Sweeney 2001), which all focus on the speech symptoms associated with velopharyngeal function. The Eurocleft Speech Group (1994, 2000) developed a cross-linguistic detailed phonetic analysis of speech for children beyond the speech development phase. Hirschberg and Van Demark (1997) and Hennisson and Hutters (1997) each proposed approaches to be adopted by the international community. The GOS.SP.ASS framework for speech assessment was developed as a comprehensive and standardized approach to assessing speech in the clinical setting in UK cleft centres (Sell et al. 1999). Closely aligned to GOS.SP.ASS, with a common set of sentence elicitation material, the Cleft Audit Protocol for Speech (CAPS) was developed specifically for audit purposes (Harding et al. 1997). Despite all these approaches, most have not been fully evaluated for their reliability and validity, although work on a UK audit tool is in progress to address this (John et al. 2003).

Rating scales of articulation have been popular. Rating scales have included categorical description (Philips and Bzoch 1969, Subtelny et al. 1972, McWilliams et al. 1990) and psychological scaling procedures (Hess 1976, Moller and Starr 1984). However, they do have limitations as an approach to measuring consonant production. For example, Hirschberg and Van Demark (1997) suggested that glottal and pharyngeal categories should be rated on a scale from one to five, with moderate, severe and very severe all as separate categories. However, there is no further definition of these categories and hence it is not clear what the differences are between them. It is frequently the case that scalar points are not defined, and it is recognized that rating scales fail to provide information on speech patterns and how improvement occurs.
Certainly more contemporary thinking emphasizes the need for a return to the profession's phonetic roots (Grunwell et al. 1993). Detailed narrow phonetic transcription should be the gold standard for this type of disordered speech. This, however, is apparently not common practice in the published literature (Lohmander and Olsson 2004). Henningsson and Hutters (1997) described the need for 'standardized speech error categories in order to facilitate simple reductions and pooling of data'. Based on phonetic transcription Harding et al. (1997) proposed categories known as speech cleft type characteristics (figure 1). Importantly, such descriptors all meaningfully convey the sound of the speech error, in sharp contrast to a traditional framework of substitution, distortion and omission. Harding et al. (1997) further grouped the speech cleft type characteristics into speech summary patterns (figure 2). These were mainly defined according to place of articulation i.e. anterior, posterior, non-oral, with a further category of passive errors for describing manner errors. For a further description of this approach, see Sell et al. (1999).

However, when data are summarized, other challenges appear. Figure 3 shows the individual categories of compensatory articulation proposed by Trost (1981), including glottal, pharyngeal, velopharyngeal frication and palatal, all shown as a dashed line. Often, however, when data are summarized using Trost's approach, the detail of these different phonetic groups is reported as one group of compensatory articulations (figure 4), limiting the possibility of comparing datasets. For example, the incidence of glottal articulation in the UCLP group in Peterson-Falzone's (1995) retrospective study of 132 UCLP children is not possible to determine for this

![Figure 1. UK system: Speech cleft type characteristics.](image-url)
Figure 2. UK system: Speech pattern summary based on speech cleft type characteristics (CTCs).

Figure 3. Trost 1981: Glottal, pharyngeal, velopharyngeal frication, backing and palatal categories.
reason. Another problem of interpretation is seen in figures 5 and 6: the Swedish system (Lohmander-Agerskov et al. 1996) is represented by a thin line and the UK system by a dotted line. The groupings appear very similar, but in the Swedish system, palatal errors are grouped as retracted errors (figure 5), whereas in the UK system, palatal errors are grouped as anterior errors. In figure 6, it is possible to see that palatal errors may either be summarized as anterior errors (UK system), retracted errors (Swedish system) or compensatory (US system), therefore making comparisons of the outcomes of these groups meaningless. There is, therefore, a need to ensure that when summarizing detail the building blocks are available to the reader.

Summary outcome measure of speech

Henningsson and Hutters (1997) stated that ‘if a simple and reliable parameter to measure intelligibility could be established it should be included in the reporting of speech in this condition’. The need for this was underlined in the UK National Clinical Standards Advisory study (Sell et al. 2001). Speech was one of eight outcomes of a cohort of children with unilateral CLP and there was strong pressure from the other disciplines in the research team for a summary outcome measure of speech. Intelligibility arguably provides this as it is considered a general judgement of speech in relation to how well a listener understands a speaker. Intelligibility, however, is controversial. Dalston et al. (1988) recommended that intelligibility and articulation should always be reported, yet others cautioned against this (Witzel
Figure 5. UK (dotted line) and Swedish system (thin line).

Figure 6. Palatal Errors summarised in different ways: US system: dashed line; UK system: dotted line; Swedish system: thin line.
They also stated how it is well recognized that intelligibility is influenced by many variables other than the speech characteristics being assessed. For example, phoneme content, stress, accent, intonation and rate are all speaker variables influencing intelligibility. Listeners’ variables include differences in their ability to resolve the ambiguities heard depending, for example, on how well the listener knows the speaker as well as his/her experience of speech disorders. External factors such as context, the message content, and background noise may also play a part. Witzel (1991) recommended that intelligibility alone should not be used to report speech results but should be used in conjunction with detailed descriptions of different speech parameters, particularly consonant production and nasality. Like Moller and Starr (1984) and Karling et al. (1993), Sell et al. (2001) reported that intelligibility ratings correlated well with the number of consonant errors, when rated by speech and language therapists who had undergone training, were aware of local accent and dialectal confounders, and were judging a controlled speech sample under controlled conditions. They cautioned that intelligibility is not suitable for reporting outcome in the developing preschool child, and should never be reported in isolation. For an in-depth review, see Whitehill (2002). Whitehill underlined the need for a global measure of speech performance, such as speech intelligibility, and indeed reported an increase in the number of studies using intelligibility as an outcome measure in investigations of speakers with cleft palate. She urged the use of appropriate methodologies in using it as an outcome measure, particularly in research studies. She argued that interval scaling is not a valid measure of speech intelligibility, and that transcription tasks, multiple choice tasks and direct magnitude estimation are more appropriate procedures. Notwithstanding, such methodologies are very challenging outside of the research setting. Subtelny et al. (1972) concluded that this was why clinicians continued to rate intelligibility rather than measure intelligibility using those methods advocated by Whitehill (2002).

**Reliability and agreement**

Frequently the reliability of perceptual speech assessment tools has not been addressed, despite recommendations to the contrary (D’Antonio and Scherer 1995, Wyatt et al. 1996, Lohmander and Olsson 2004). Hayden and Klimacka (2000) stressed how important inter-rater reliability is with the recent wide recognition of the need for multi-centre studies. Despite all the above aspirations for detailed phonetic analysis, one of the continuing challenges is that complex speech disorders are often associated with low transcriber agreement (Shriberg and Lof 1991). More detailed transcription provides greater information but has limited reliability (Shriberg and Lof 1991, Kent 1999). Howard and Heselwood (2002) summarized the phonetic features that have been particularly problematic, many of which are associated with disordered speech and the cleft palate condition. Gooch et al. (2001) reported the difficulty associated with transcription of compensatory articulation. Gibbon and Crampin (2001) perhaps captured something of why this should be in their statement ‘the activity of transcription affords an indirect representation of the actions of the articulators, with the result that articulatory information must be inferred by the transcriber from an accumulation of complex clues contained in the
acoustic signal’. This may account for why even under optimal listening conditions, phonetic transcription data may not agree with physiologic or acoustic records (Shriberg et al. 1984).

Gooch et al. (2001) recently examined the reliability of the phonetic transcription of compensatory articulations by two groups of ten speech–language pathologists based on digitized audio recordings. Group 1 included speech–language pathologists who were experienced in evaluating children with cleft palate, and group 2 included speech–language pathologists who were not. They found that on average only 40% of transcriptions agreed with the ‘gold standard’, i.e. an expert’s transcriptions. Like Keuning et al. (1999), they reported that experienced listeners did better than inexperienced listeners but not as well as predicted, and furthermore that the errors categories judged to be most easy to transcribe had the lowest reliability. Studies have often reported the experience of two listeners, with an accumulated large number of years working in the field, as a measure of listener reliability. Kent (1999) raised doubts about the degree to which experience alone guarantees satisfactory levels of inter-judge agreement, and suggested that it is more likely that colleagues who have worked together come to listen in a similar way. The role of experience is currently ill understood but it is inappropriate that this factor alone should be considered as an index of reliability. Furthermore, listener reliability established over many years may well be based on outdated assessments. Keuning et al. (1999) also found that experience with cleft palate speech did not guarantee higher intra-judge reliability.

Gooch et al. (2001) also highlighted the difference between identifying an error and transcribing it. They stated that the lack of confidence some listeners had in their transcription abilities could be attributed to a lack of continuing professional development in the field of transcription workshops within the three years before the study. Kreiman et al. (1993) reported similar challenges in the perceptual evaluation of voice quality, and attributed variability in ratings to several sources, including listeners’ backgrounds and biases, the task used to gather the ratings, interactions between listeners and tasks, and random error.

Instrumentation has been very useful in perhaps helping understand why reliability might be so challenging in the cleft palate field. Electropalatography (EPG) has shown that there is variability of tongue placement according to the anticipated target, so the tongue is more anterior in anticipation of /t/ targets and more posterior in anticipation of /k/ targets, and that palatal errors may involve lateral release and lateral friction. Santelmann et al. (1999) and Gooch et al. (2001) have both questioned the reliability of listeners for this error type, using English samples and listeners whose first language is English. This is because the target sound lays within the bounds of phonemic categories for the listener’s own language, that of both alveolar and velar plosives (Eurocleft Speech Study 1994, 2000, Santelmann et al. 1999). EPG has also provided evidence for double articulations, where there are two simultaneous places of articulation co-occurring. This validates the findings of Gooch et al. (2001) who reported that glottal double articulation had a profound effect on lowering listener reliability. Reliability is probably even more challenging in cross-linguistic studies as found in the Eurocleft Study (1994, 2000), which reported that the alveolar consonants, which are the most vulnerable consonants in the cleft palate condition, were the most difficult targets on which to achieve listener agreement. Instrumentation, using nasometry, has also defined important differences in nasality across languages and dialects which will

Rating scales, using descriptive category judgements and equal interval scaling, remain the favoured approach for dealing with resonance disorders and nasal airflow errors. Direct magnitude estimation and paired comparison have also been used (Subtelny et al. 1972, Moller and Starr 1984, McWilliams et al. 1990). Recently Whitehill et al. (2002) concluded that equal interval scaling may not be a valid method for the evaluation of hypernasality, and recommended the use of direct magnitude estimation in research studies. They also acknowledged, however, that due to practical considerations, equal interval scaling should remain the approach in the clinical setting. Riski (2001) suggested that reliability was improved by reducing the number of choices on a rating scale. Frequently rating scales are of different lengths, so for example, three-, five-, seven- and even ten-point scales have been used, not only limiting the ease of comparing outcomes, but also increasing the likelihood of poor reliability on the longer scales. The use of the mean of multiple listener judgements has been popular (Moller and Starr 1984, Starr et al. 1984 Counihan and Cullinan 1970), although Kreiman et al. (1993) pointed out that when using data averaged across clinicians, important aspects of an individual’s perceptual behaviour may be lost. One approach to reducing measurement error is the use of consensus listening particularly when one is agreeing narrow phonetic transcription (Shriberg et al. 1984), although its time-consuming nature is acknowledged. They advised that attention is paid to avoiding bias, such as being aware of the effect of the higher ranked or more forceful transcriber. Another potential source of variability in methodology is the number of times a tape can be listened to. Amorosa et al. (1985) advocated listening as often as necessary, but Shriberg and Kwiatkowski (1980) suggested that listening should be restricted to no more than three times. Practical factors may again be an issue for consideration in this debate. Howard and Heselwood (2002) also discussed consensus listening, and concluded that ignoring or excluding data because it is difficult to transcribe, is not an appropriate strategy.

Active structured training has been advocated over several training sessions with external reference taped samples, rather than depending on the internal standard of the experienced judge (Young 1969, McWilliams and Philips 1979, Kent 1999, Keuning et al. 1999, Gooch et al. 2001). Trost (1981) advocated that examiner expectancy should also be reported, such as the amount of detail listeners are given about the cohort under study. For example, listeners who are informed they are analysing speakers with 22q11 deletion syndrome may be biased towards glottal and pharyngeal type errors in their transcription. John et al. (2003) advocate a methodology that uses consensus listening both for ongoing training requirements combined with providing outcomes for audit studies. This may be a practical solution to the time-consuming tasks of analysis and the requirement for continuing professional development.

Another probable source of poor reliability is ill-defined terminology. Kent (1999) discussed the problems of an equivalent understanding of terminology and their definitions. This is an important issue in the CLP field. Whitehill (2002), in a review of intelligibility in cleft palate, reported confusion over terminology in more than 20% of the 56 studies reviewed. Table 1 lists the wide range of descriptors used
for nasal airflow. All these terms are in use and many are overlapping, adding to confusion. An important step forward is to define and agree terminology and definitions (Kuehn et al. 2002, John et al. 2003).

Who should decide the outcome?

Several authors have raised the issue as to who is the most appropriate judge of speech (Starr and Moller 1984, Bagnall and David 1988, Riski 1996, Witt et al. 1996, 1997, Keuning et al. 1999). One recurring trend which has emerged in the literature, is the suggestion that the team’s speech and language therapist is ‘too critical’, whereas untrained listeners (such as non-specialist therapists, other Cleft Team members, parents of affected or non-affected children, affected children or their normal peers) may all add real-life significance to clinical speech assessments. Bagnall and David (1988) reported a study in which naïve listeners ranked the speech of 21 school aged cleft children and their peer group. They found that the cleft children spoke ‘less acceptably’ than their peers, and concluded that naïve listeners were a useful adjunct to the traditional assessment by the speech and language therapist. The children were, however, still undergoing management and included those that had required therapy. Furthermore, the speech sample depended on the child’s expressive language skills, and the term ‘acceptable speech’ was not well defined. Witt et al. (1996) concluded that peer group evaluations defined the morbidity of cleft palate speech in terms that are most relevant to the patients themselves. They also suggested that this type of listener might safeguard against offering treatment that may be unnecessary. Riski (1996), in an eloquent critique of Witt et al., concluded that the expert and lay scales used were not valid assessments of the speech of speakers with cleft palate, that they failed to demonstrate the reliability of some of the expert and lay listeners, and it was not possible to compare the results of the two groups of listeners using the expert and lay scales. Although it is possible that lay listeners may theoretically add validity to outcomes, presently the methodology has not been adequately developed. Furthermore, such assessments should never replace the expert listener’s assessment but should serve as an adjunct to it. Kuehn and Moller (2000) advocated that outcome measures for speech must include both professional and patient and/or family judgements (ACPCA 1993).

| Table 1. Examples of a wide range of descriptors for nasal airflow consonant errors |
|---------------------------------------------|------------------|
| Phoneme-specific nasality                  |                  |
| Phoneme-specific nasal emission            |                  |
| Phoneme-specific nasal turbulence          |                  |
| Active nasal fricative                     |                  |
| Passive nasal fricative                    |                  |
| Velopharyngeal fricative                   |                  |
| Nasopharyngeal fricative                   |                  |
| Nasal snort                                |                  |
| Nasopharyngeal snort                       |                  |
| Nasal lisp                                 |                  |
| Nasal rustle                               |                  |

Debbie Sell
Probably, however, a more important issue to be explicit about is who the expert listener should be. Wyatt et al.’s (1996) proposal for independent assessment by specialist speech and language therapists allows for bias to be controlled. Mehendale and Sommerlad (2003) considered that blind consensus listening, and considered it ensured a consistent approach to assessment. They acknowledged its disadvantages, including the lack of direct interaction between the therapist and the patient, the loss of visual clues, and the potential reduction of sound quality in recordings compared with live assessment, but nevertheless concluded its advantages far outweighed the disadvantages, particularly that of bias.

Keuning et al. (1999) found that specialist speech pathologists were generally more reliable than non-specialists, although the effect was small. Overall, they found wide variation in listener judgements of speech samples, but unfortunately, they did not include any training or consensus listening training. This has generally been a weakness of several of the methodological studies reviewed here (Gooch et al. 2001). In addition, Santelmann et al. (1999) used 20 listeners, ten of whom were untrained and ten of whom were trained in general phonetics. John et al. (2003), in contrast, reported positive outcomes following a period of training for specialist speech and language therapists. Future studies aimed at investigating appropriate methodologies should be based on listeners who are specialist therapists and should include a period of training.

**Impairment, activity, and participation**

The focus of outcome studies in cleft palate is almost exclusively at the level of the impairment, for example the dysfunction as a result of pathology. The World Health Organization (1980, 2001) advocated describing the impact of impairment on an individual’s functional performance (activity) and their social consequences (participation). In other words, there is a need to measure the impact of the speech impairment on a patient’s social functioning. Enderby et al. (2000) reported that all three dimensions (impairment, activity and participation) were affected in this client group. The need for an outcome measure to capture such information was clearly seen in a study on palate re-repair (Sommerlad et al. 2002). They described patients who had low impairment scores on the perceptual scales but high scores on disability/activity, with observations such as worse when unwell and/or tired, restricted talking in front of a large group of people, and avoidance of speaking on the telephone. In other words, their speech disorders had significant limiting consequences disadvantaging them socially. Currently such issues are not routinely captured in outcome measures, which focus on impairment. This is an area of work that needs to be developed in this field.

**Conclusion**

This review attempted to bring together issues involved in the perceptual measurement of speech associated with cleft palate and related disorders. Measuring at the impairment level involves the highly skilled and complex process of data sampling, data collection/recording, archiving, a standardized approach to data analysis and interpretation. It is recommended that outcome measures need to be extended to incorporate more functional issues, which affect quality of life. The
need to meaningfully capture the sound of speech based on detailed phonetic analysis is restated, and suggestions are made about how reliability may be addressed. It has emerged that, based on currently available methodologies, specialist therapists should be the primary arbiters of outcome. In studies determining methodologies, such therapists should be included and training is suggested as an essential component in this. Care should be taken in reporting outcomes based on clinical data. Blind independent analysis of speech data by specialist therapists should be the gold standard methodological approach when reporting audit and research outcomes.

Acknowledgements

This paper is based on the keynote presentation, Unravelling Speech Assessment in Cleft Palate, given at the 9th International Congress on Cleft Palate and Related Craniofacial Anomalies, Gothenburg, Sweden. The author completed the paper during a Distinguished Visiting Fellowship at the Institute of Advanced Studies, La Trobe University, Melbourne, Australia.

References

AMERICAN CLEFT PALATE CRANIOFACIAL ASSOCIATION, 1993, Parameters for Evaluation and Treatment of Patients with Cleft Lip and Palate or Other Craniofacial Anomalies (Chapel Hill, NC: ACPCA).


HARDING, A., HARLAND, K. and RAZZEL, R., 1997, *Cleft Audit Protocol for Speech (CAPS)* (available from Speech/Language Therapy Department, St Andrew’s Plastic Surgery Centre, Broomfield, Chelmsford, Essex).


