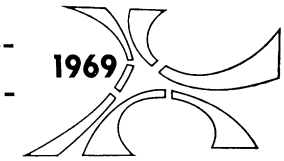
A Study of Speech Improvement Following Palatopharyngeal Flap Surgery



KERSTIN ENGSTRÖM, M.S., Logoped
BJÖRN FRITZELL, M.D.
BENGT JOHANSON, M.D.
Göteborg, Sweden

The pharyngeal flap is a frequently used procedure for improving velopharyngeal competency. In order to properly evaluate the results of such procedures it would seem necessary to make a comparison between the patient's performance before and after the treatment. Several different procedures may be used for examining velopharyngeal competency, such as oral air pressure ratios and cinefluorographic studies. However, since the major concern in connection with velopharyngeal incompetency is its effect on speech, an evaluation of the results of pharyngeal flap operations should in some way include an evaluation of speech quality.

Unfortunately speech quality is not easily measured as it comprises many variables. Rating scales have been used for the evaluation of such variables as hypernasal quality and articulation defectiveness. Equal-appearing-intervals scales have most frequently been used (*9*). Rating scales have sometimes been considered too subjective as they require a judgment on the part of the listener. Although efforts have been made to use more "objective" methods such as spectrographic analysis, such methods have not found common use. Moll (*5*) pointed out that since the purpose of speech is communication, it is not logical to eliminate the listener from the evaluation procedure. It would thus be desirable to find a method which would provide relatively objective measurements without excluding the important role of the listener as the receiver in the communication process. In intelligibility tests listeners are used without their making any subjective judgments of what they hear. They simply write down the words, syllables, or phrases presented to

Miss Engström, formerly research assistant at the Department of Otolaryngology, University of Göteborg, Sweden, is now clinical supervisor at the Speech and Hearing Center, University of New Mexico. Dr. Fritzell is head of the program of phoniatrics of the Department of Otolaryngology, and Dr. Johanson is professor and head of the Department of Plastic Surgery, University of Göteborg, Sweden.

This study was supported in part by the Swedish Medical Research Council Grant ※17x-756-01.

This paper was presented at the 1969 International Congress on Cleft Palate, Houston.

them as they understand them. The proportion of correctly understood items gives a measurement of the intelligibility of the speech.

Intelligibility tests have been used in a few studies concerned with cleft palate speakers for comparing the speech quality of the same speaker under different conditions. Prins and Bloomer (*6*) used intelligibility tests to analyze the results of speech therapy for 10 patients. Subtelny, Sakuda, and Subtelny (*8*) used the same procedure to compare the speech of 23 patients with and without the use of speech obturators. Shupe, Subtelny, and Wilson (*7*) used intelligibility tests to compare speech before and after pharyngeal flap surgery.

The present study developed from the need for an objective evaluation of the effects of pharyngeal flaps upon speech. The purpose was to obtain a quantitative measurement of the change in speech as a part of the clinical evaluation of each patient. Intelligibility tests were chosen as they were felt to offer the desired objectivity. A preliminary study has been described by Fritzell (*2*). The results of the intelligibility tests were compared with results of subjective listener judgments of changes in speech quality.

Procedures

Subjects. From January 1961 to April 1965, 72 patients, ten years of age and older, underwent pharyngeal flap surgery at the University Hospital in Göteborg, Sweden. Each of these patients was seen for a speech evaluation at the Phoniatric Department of the same hospital prior to surgery. A follow-up speech evaluation was made one to four years after surgery for 58 of the patients. This study is based on speech recordings of 53 of these patients and 15 additional patients who underwent surgery before 1961. Five patients of the 58 had to be eliminated from the study because of difficulties with the recording equipment.

The ages of the patients ranged from 10 to 54 years with a mean age of 25 and a median age of 21. Different types of clefts, as well as a few cases of velar paralysis, were included in the study. Most patients with palatal clefts had undergone primary repair procedures several years prior to the pharyngeal flap surgery. Thirteen patients used obturators before this surgery. The surgical procedure in all cases consisted of superiorly based pharyngeal flaps usually combined with V-Y push-back palatoplasty (Wardill-Kilner).

Recordings. Speech recordings were made at the time of the preoperative and the follow-up speech evaluations. For twenty patients, recordings were also made approximately one month following surgery. Although two sets of recording equipment were used, the same equipment was always used for the initial and repeat recordings of a given patient. The equipment used for recording, editing, and playback

of the material upon which this study was based has been described in a previous article (*3*).

INTELLIGIBILITY TESTS. Forty-one patients were included in this part of the study. Eight phonetically balanced lists of 50 monosyllabic Swedish words were used. The lists were chosen from among those used for discrimination testing in speech audiometry (*1, 4*). At the time of each evaluation, the patient was recorded while reading one list at a rate of one word every five seconds. The same list was never used more than once with a given patient. Preoperative recordings for subjects with obturators were made with the obturators in place.

Six listeners were used to establish intelligibility scores for a given patient. The order of presentation of the lists was rotated among the listeners to eliminate any adaptation effect on the part of the listeners. The listeners were instructed to interpret each utterance as a one-syllable Swedish word and to write down the word they heard. The mean percentage of words correctly understood by the six listeners was expressed as a percentage score of intelligibility. Separate intelligibility scores were obtained for preoperative, postoperative, and follow-up recordings and the difference provided a measurement of speech change.

Each listener listened to one, two, or three patients consecutively, but each word list only occurred once in a listening sequence. Some listeners listened to two or three different sequences. A minimum period of two weeks was then required between listening tasks. In order for the listeners to learn the task, a standard list was presented first in each listening sequence. This list was different from the eight used in the study, but was chosen from the same source. It was read by a patient with repaired cleft palate who was not included in the study and whose speech intelligibility was relatively low.

The listeners differed with respect to vocational and educational background. No systematic differences were found among the results obtained by different categories of listeners. All listeners were required to have a history of good hearing. However, it was not possible to provide hearing tests for the listeners. The recordings were played back through earphones at a comfortable listening level.

Ten cleft palate patients with varying levels of speech intelligibility were used as a control group. Each of these patients was recorded on eight occasions with a minimum time lapse of one week between recordings. A different word list was recorded each time so that each patient read each of the eight lists once. No treatment was given to these patients between recordings.

A statistical analysis revealed no significant differences among the eight lists. It was thus assumed that for the purposes of this study the lists were of equivalent intelligibility.

In order to determine what difference in intelligibility between two

TABLE 1. Standard error and minimum significant differences between intelligibility scores obtained from repeated recordings of the same patient. The analysis was performed on data from ten control subjects.

average intelligibility	S_e	*significant differences between intelligibility scores*	
		$P = .05$	$P = .01$
0–100% (10 patients)	3.16	7	9
0–80% (5 patients)	3.80	8	10
80–100% (5 patients)	2.34	5	7

recordings would be significant, an analysis of variance was performed on the data from the ten control subjects. The standard error of the individual measurements was computed for the total sample as well as for two sub-samples derived according to average intelligibility. The results are shown in Table 1. The variability was found to be greater for patients with lower average intelligibility. Thus a difference between recordings of 10% would be required for patients with preoperative intelligibility scores below 80%, in order to show significant change in speech intelligibility at the .01 level of confidence. For patients with preoperative scores above 80, a difference of 7% would be significant at the .01 level of confidence.

LISTENER JUDGMENTS. Sixty-six patients were included in this part of the study. Thirty-nine of these were also evaluated with intelligibility tests. Three short sentences were used. For each patient, the sentences from the preoperative recording were paired with the same sentences from the follow-up recording so that the listeners would hear the same sentence from the two recording times in direct succession. The three-sentence pairs for each patient were presented one after another, all in the same order. The order was changed randomly among the patients so that the listeners would not know which was the follow-up recording.

The listener tasks were a) to decide with respect to each patient whether the "best" recording was presented first or last in the three-sentence pairs, and b) to rate the perceived difference in speech quality according to three categories: "hardly noticeable," "clear," and "very clear". The task was preceded by careful instructions and a training series of 10 patients judged by the experimenter as representing various degrees of speech change and levels of speech quality.

Eight listeners were used: one phoniatrist and three speech clinicians from the Department of Phoniatrics, and two surgeons and two nurses from the Department of Plastic Surgery. The recordings were presented to four listeners at a time through a loudspeaker which was placed at approximately the same distance from all listeners.

The procedure provided a rating of the speech change in three

degrees. The change could be positive or negative, depending on whether the listeners had chosen the follow-up recording as the better one. The values +5, +15 and +25 were assigned to the three degrees of positive change, and the values −5, −15 and −25 were assigned to the three degrees of negative change. The median of the eight listener judgments was chosen as the value for each patient.

Each patient was thus given a median rating of speech change between the values −25 and +25. The obtained values were divided into five classes representing two degrees of deterioration, no observed change, and two degrees of improvement. There was good agreement among listeners. For 62 of the 66 patients the judgments of six or more listeners differed by no more than one category. For two patients there was total agreement among all eight listeners as to degree as well as direction of change. Five or more of the listeners agreed completely in forty cases.

Results

INTELLIGIBILITY TESTS. In Figure 1, intelligibility scores obtained from recordings made at the final follow-up evaluation are plotted against preoperative scores. Each patient is represented by a circle or a dot, the latter showing patients who used obturators before the operation. Marks above the diagonal indicate increased intelligibility after the treatment. Although the increase is not significant in all cases, the figure shows that all but four patients obtained a higher score after treatment.

Figure 1 shows clearly a serious limitation of the method used. The upper horizontal line represents a ceiling above which no further improvement can be demonstrated since intelligibility scores above 100%

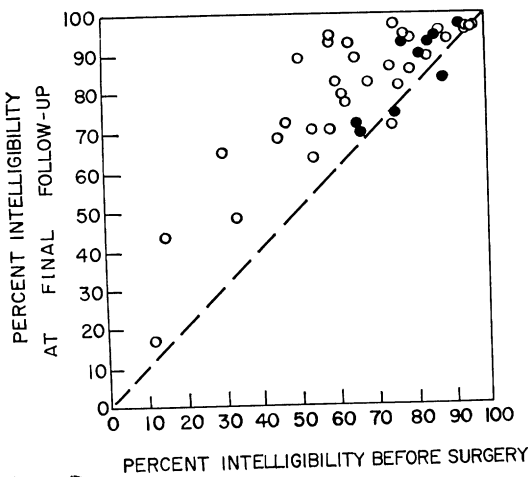


FIGURE 1. Results of listener judgments (41 patients). The filled circles represent patients who used obturators prior to the pharyngeal flap operation. Marks above the diagonal indicate increased intelligibility.

TABLE 2. Results of intelligibility tests for patients with different levels of preoperative speech intelligibility and for the total sample.

change in intelligibility	*% intelligibility before surgery*				*total*
	12–50	*51–65*	*66–80*	*81–96*	
signif. decrease	0	0	0	0	0
no signif. change	1	0	6	9	16
signif. increase	1	5	5	3	14
marked increase	4	5	2	0	11
no. of patients	6	10	13	12	41

are impossible. Thus, this method is less able to reveal improvement in patients whose preoperative intelligibility scores were relatively high.

Table 2 shows the results of the intelligibility tests among patients with different levels of preoperative speech intelligibility. A change in intelligibility in the −6 to +6% range was considered to be nonsignificant. The classification "significant increase" includes changes of +7 to +17%. Three patients with preoperative intelligibility scores of 85 to 88% showed improvement of 7%. This was significant at the .01 confidence level. One patient with a preoperative score of 54% showed improvement of 9%, significant at the .05 confidence level. All other patients reported in this group showed improvement of 11% or more. The classification "marked increase" includes changes of +21 to +37%.

No patients were found to have significantly lower speech intelligibility following surgery. Significantly increased intelligibility was found for 25 of 41 patients, or slightly less than two-thirds of the total sample. If those patients are excluded whose preoperative scores were over 80%, significantly increased intelligibility was found for 22 of 29 patients. Considering only patients with preoperative intelligibility scores of 65% or less, improvement was found for 15 of 16 patients.

For 20 patients, postoperative recordings of word lists were made approximately one month following the surgery in addition to the preoperative and the follow-up recordings. Five of these patients had preoperative intelligibility scores above 80%; none of these five showed a significant change in intelligibility at either the immediate postoperative or the follow-up recording. The results for the 15 patients whose preoperative scores were 80% or lower are shown in Figure 2. A difference in intelligibility of 10% or more was considered significant (.01 level of confidence). Only two patients showed significant increase in intelligibility approximately one month after surgery. These two patients retained the same degree of improvement at the follow-up recording. The majority of those patients whose follow-up recordings showed improvement seemed to have made the major part of their improvement after the first month following surgery. Most patients, and

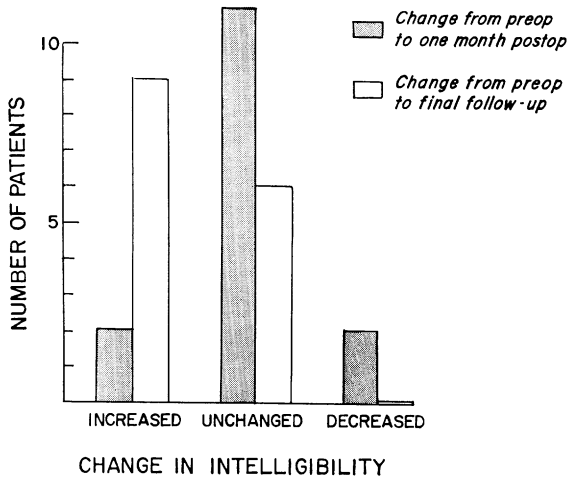


FIGURE 2. Change in intelligibility from the preoperative recording to each of two postoperative recordings, made approximately one month after the surgery and one to four years later (15 patients).

all of those who improved significantly between the first postoperative and the follow-up recordings, received speech therapy following the surgery. In some cases, however, this therapy amounted to only a few sessions. No conclusion can be drawn from this study as to the effect of speech therapy upon the final results.

LISTENER JUDGMENTS. The result of the listener judgments are presented in Figure 3. The category "much improved" includes median ratings of +16 and higher; "improved" includes ratings of +6 to +15; "unchanged" includes ratings of −5 to +5; "worse" includes ratings of −15 to −6; "much worse" includes ratings of −25 to −16. Two-thirds of the total sample, or 24 patients, were rated as improved and one-third as unchanged or worse. The two patients whose speech was rated as "much worse" and one of the three patients whose speech was rated as "worse" used an obturator before surgery. The distribution of speech ratings of the 39 patients studied with both intelligibility tests and listener judgments was similar to that of all 66 patients.

COMBINED RESULTS OF THE TWO METHODS. The listener judgments as well as the intelligibility tests showed that approximately two-thirds of the total number of patients studied had significantly improved their speech following the surgery. Except for the fact that the listener judgments showed a few cases of deterioration not demonstrated by the intelligibility tests, it would seem as though the two methods had given similar results. Important differences were found, however, when the results of the two methods were compared for individual subjects. In Figure 4 the results of listener judgments are plotted against the results of the intelligibility tests for the 39 patients who were included in both parts of the study.

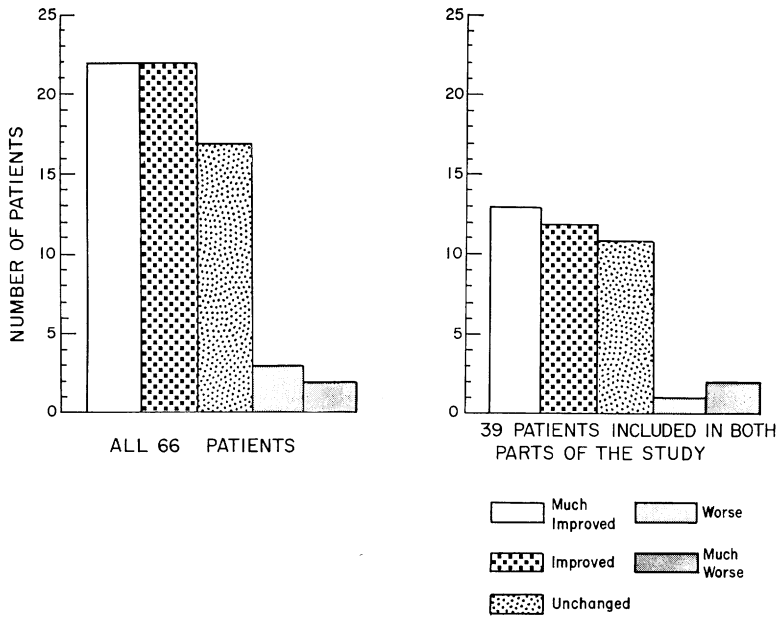


FIGURE 3. Results of listener judgments (66 patients).

Different marks have been used for four different categories of preoperative intelligibility. Within each category lines are connecting the median values of intelligibility change for patients with the same degree of speech change according to listener judgments. The best correlation between the results of the two methods was found for patients with preoperative intelligibility scores in the range between 50 and 80%. Essentially no correlation was found for patients with preoperative scores above 80%. Although little or no improvement was demonstrated with intelligibility tests for these patients, the listener judgments showed a distribution of speech change which ranged from "much worse" to "much improved". Patients who had a preoperative intelligibility score below 50% showed limited or no improvement as measured by listener judgments. Here the intelligibility tests gave a distribution of change ranging from +5% to +35%. In no case were the results directly opposed in such a way that a patient's speech was found to have improved by one method and deteriorated by the other. However, of 15 patients whose speech showed no significant change according to intelligibility tests, three were judged as worse and five as improved by listener judgments. Of ten patients whose speech was considered unchanged by listener judgments, four were found to be significantly improved according to the intelligibility tests.

Combining the results of both methods, and assuming that a change in speech is significant if demonstrated by at least one method, 29 of the 39 patients, or approximately 75%, were found to have improved; six

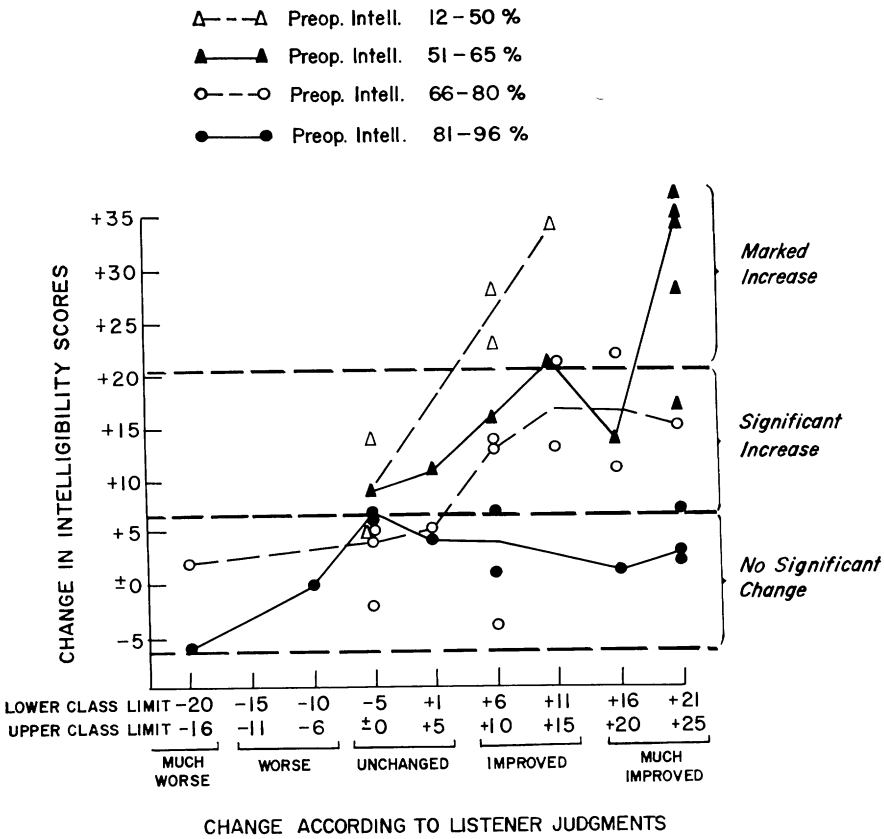


FIGURE 4. Combined results of intelligibility tests and listener judgments (39 patients). The curves are connecting the median values of change in intelligibility for patients with the same degree of speech change according to listener judgments and the same range of preoperative intelligibility.

were unchanged and four were worse. Of 30 patients who did not use an obturator prior to surgery, 26 patients, or 87%, were found to have improved; three were unchanged and one was worse. Patients who used obturators before the operation showed somewhat different results; three had improved their speech, four were unchanged, and two were worse than when they used their obturators.

Discussion

From the results of the intelligibility tests shown in Table 2 it would appear that those patients with the lowest speech intelligibility would profit most from the pharyngeal flap. Among patients with preoperative scores of 65% or less, improvement was demonstrated for all but one. Only approximately half of the patients with preoperative scores of 66 to 80% and one fourth of those with preoperative scores above 80% were shown to have improved significantly. The limitation of the method

caused by the ceiling at 100% has already been pointed out. Although this ceiling limits the amount of improvement which can be demonstrated, it would not be expected to limit the extent to which a significant increase in intelligibility could be shown when preoperative scores are 80% or lower.

Whereas the results of the intelligibility tests thus would indicate that patients with a more severe speech deficiency more frequently benefit from a pharyngeal flap, the results of the listener judgments seem to at least partly contradict this. The method of listener judgments used in this study made no reference to preoperative level of speech quality. The intelligibility scores are therefore the only available indications of preoperative speech quality. The combined results of the two methods (Figure 4) show that listener judgments demonstrated speech improvement in approximately half of the patients with preoperative intelligibility scores above 80%, and in almost two-thirds of the patients with scores between 66 and 80%. Thus listener judgments demonstrated improvement in a larger proportion of patients with preoperative scores above 65% than did intelligibility tests. Listener judgments also demonstrated deterioration of speech in three patients in this category. Listener judgments thus appear to be more sensitive than intelligibility tests to changes in speech quality in relatively "good" speakers. The opposite appears to be true for the relatively "poor" speakers. For these speakers changes in speech quality which caused an increase in speech intelligibility of 20% or more were reflected only as small changes in listener ratings of subjectively perceived speech quality differences. It may be speculated on the basis of these results that listeners are more sensitive to differences in speech quality as this quality approaches "normal". Even though our listeners were instructed to judge only the degree of difference between paired recordings, it appears as though their judgments were influenced by whether the best recording came close to normal.

The apparent low validity of intelligibility tests for measuring speech improvement in speakers with relatively high intelligibility might be improved if the total range of intelligibility scores could be lowered. An attempt was made to accomplish this by lowering the playback intensity. This resulted in increased variability among lists read by the same speaker without intervening treatment and in an increased number of atypical errors. This procedure was therefore discarded. It is possible, however, that the method could be made somewhat more effective by a reconstruction of the word lists. The lists used were phonemically balanced to represent the language in general. Some words were frequently misunderstood by the listeners and others were always correctly understood. The lists thus contained many words which discriminated poorly between good and poor speakers. These words could be eliminated and

others added which are more sensitive to the particular errors which are typical for the speech problem studied.

Another limitation was noted in the use of intelligibility tests. In order for the results to be valid, the speaker must read each word correctly. Several patients had to be eliminated from the study because of poor reading ability. Although reading errors were generally of a different type than errors caused by the speech disorder, it was sometimes difficult or impossible for the experimenter to determine whether the patient was reading correctly. If reading errors were undetected by the experimenter, the listener would hear an incorrect word and intelligibility scores would be affected. This problem may have been partially solved by presenting the words auditorily as well as visually to the patients. This was not done because of the high incidence of hearing loss among these patients.

Finally, intelligibility tests as used in this study are not practical as a clinical tool, as they are extremely time consuming. The listener judgments are easier to carry out, and have the advantage that the same listeners may be used for all patients, whereas the intelligibility tests require constant renewal of the listeners.

The method of listener judgments used in this study was felt to give a fairly objective reference point on the scale as the listeners were required to choose one of two recordings as the better. There was, however, no way of controlling whether the steps on the scale, provided by the listeners' ratings of the difference, were equidistant. A disadvantage of the method appeared to be the tendency for the listeners to evaluate the speech change more according to how nearly normal the "best" recording was, than according to the actual difference between the recordings. If the method is to be used as a means of measuring speech change in patients with all degrees of speech proficiency, it would therefore seem necessary to incorporate a classification of the speech quality before treatment.

Although no attempt was made in this study to determine whether the patients attained "normal" speech after treatment, our impression was that such a perfect outcome occurred in few if any cases. The majority of the patients appear to have improved their speech however. Of 55 patients who did not use an obturator before treatment, 44 were found to have improved their speech as indicated by one or both of the methods, and two were considered worse according to the listener judgments. In general, better results were found for patients under 21 years of age than for older patients. In view of the uncontrolled variables, no statistical analysis was performed to determine whether this difference was significant.

Out of a total of 13 patients who used obturators before surgery, four had improved their speech to some degree, five equalled the speech

results obtained through obturator use, and three were rated as having poorer speech quality following surgery than previously with their obturators. All of these patients were over 25 years of age. The primary purpose of the pharyngeal flap operation for those patients who used obturators was to free the patient from problems incident to their use. In this instance "success" could be defined as the achievement of speech which was at least as good as previously obtained with the prosthetic appliance. Several of the patients who could not be shown to have improved their speech, or whose speech was judged to be slightly worse, indicated that speaking was easier and were pleased with the results. These results would suggest good prognosis for successful replacement of an obturator with natural tissue.

The purpose of pharyngeal flap surgery is to give the patient a better tool with which to speak. After surgery has been performed, articulation habits have to be changed. Whether help in the form of speech therapy is received or not, this relearning would be expected to take some time. The 20 patients with whom three recordings were made seem to illustrate this. Only two of these improved significantly between the first and second recordings, but nine improved from the first to the third.

Summary

Intelligibility testing and listener judgments were used to evaluate the effect of pharyngeal flap operations on speech in 68 adult subjects with velopharyngeal incompetence. Preoperative speech recordings were compared with recordings obtained three to four years after the surgery. The results indicated that a) at least 80% of the patients studied who did not use obturators before the operation improved their speech following pharyngeal flap surgery, and b) two-thirds of the patients who used obturators maintained or improved their speech proficiency when the prosthetic appliance was replaced with natural tissue. A comparison between the results of the two procedures indicated that they provided complementary information about speech improvement. Whereas intelligibility tests appeared to be the more effective procedure for measuring speech change in relatively "poor" speakers, subjective listener judgments appeared more sensitive to changes in speech quality in near normal speakers.

reprints: *Kerstin Engström, Clinical Supervisor*
Speech and Hearing Center
University of New Mexico
Albuquerque, New Mexico 87106

Acknowledgment. The authors gratefully acknowledge the assistance of Erik Wallin, Ph.D., Department of Education, University of Göteborg, as statistical consultant.

References

1. FANT, G., and G. LIDEN, Swedish word material for speech audiometry and articulation test. *Acta Oto-laryng. Suppl. 116,* 189–210, 1954.
2. FRITZELL, B., Speech improvement following palatoplasty with elongated pharyngeal flap. *Folia phoniat., 12,* 118–123, 1960.
3. FRITZELL, B., and KERSTIN ENGSTRÖM, Mätningar av talförbättringen hos patienter som opererats med palatopharyngeallambå. *Tale og Stemme, 29,* 1–19, 1969.
4. LIDEN, G., Speech audiometry, an experimental and clinical study with Swedish language material. *Acta Oto-laryng. Suppl. 114,* 1954.
5. MOLL, K. L., Objective measures of nasality. *Cleft Palate J., 1,* 371–374, 1964.
6. PRINS, D., and H. H. BLOOMER, A word intelligibility approach to the study of speech change in oral patients. *Cleft Palate J., 2,* 357–368, 1965.
7. SHUPE, L. K., JOANNE D. SUBTELNY, and D. K. WILSON, Speech intelligibility measures of cleft palate speakers before and after pharyngeal flap surgery. Presented at the 44th annual convention of the American Speech and Hearing Association, Denver, 1968.
8. SUBTELNY, JOANNE D., M. SAKUDA, and J. D. SUBTELNY, Prosthetic treatment for palatopharyngeal incompetence: research and clinical implications. *Cleft Palate J., 3,* 130–158, 1966.
9. VAN DEMARK, D. R., Misarticulations and listener judgments of the speech of individuals with cleft palates. *Cleft Palate J., 1,* 232–245, 1964.