An intervention study, addressing intelligibility of tracheoesophageal speech

P. Jongmans (MA), M.A. van Rossum (SLP, PhD), C.J. van As-Brooks (SLP, PhD), M.F. Polak (SLP), P.E.M. Lansdaal (SLP), F.J.M. Hilgers (MD, PhD) & L.C.W. Pols (PhD)

Today’s session

- What is Tracheoesophageal Speech?
- Results from previous experiments
- Intervention study: importance, set up and results
- Conclusion
- Discussion

Intelligibility study: listening experiment

Speakers:
11 male Dutch TE speakers
Provox® voice prosthesis
Mean age 66.9 years (44 to 78)
Mean time after TLE 9;4 years (2;2 to 17;5)

Speech Material

- CV, VC, VCV syllables
- h-V-t syllables
- Consonant clusters in combination with vowel
- SUS
- Spontaneous speech

Online listening experiment

- Ten naïve listeners
- Online experiment developed by Rob van Son
- www.fon.hum.uva.nl/Service/Experiment/index.html
- Comfortably at work or home
- More flexible in when they listen and for how long
- Logistically much easier to arrange
Consonants

<table>
<thead>
<tr>
<th>Manner</th>
<th>initial</th>
<th>medial</th>
<th>final</th>
</tr>
</thead>
<tbody>
<tr>
<td>plosive</td>
<td>p b t d k</td>
<td>p b t d k</td>
<td>p t k</td>
</tr>
<tr>
<td>fricative</td>
<td>f v z h x j</td>
<td>f v z h x j</td>
<td>f x x</td>
</tr>
<tr>
<td>nasal</td>
<td>m n</td>
<td>m n p</td>
<td>m n</td>
</tr>
<tr>
<td>approximant</td>
<td>v j r</td>
<td>v j r</td>
<td>l r</td>
</tr>
</tbody>
</table>

General observations

- Overall percentage correct score of 75% for TE speakers (compared to overall percentage correct score of 94% for normal laryngeal speakers in CVCVC syllables as found by Pols (1983))
- Very high inter-rater reliability scores for all consonant positions (.992, .996, .998 for initial, medial and final position) allow us to use averages

Manner of articulation

- Significant differences in scores are found for the different manners of articulation
- Post Hoc test (Tukey HSD): fricatives have a significantly lower score than plosives, nasals and approximants in initial and medial position. Nasals score significantly lower in final position

Voicing confusions

- In initial position, voiceless sounds became voiced more often than voiced sounds became voiceless (chi-square, p < .01) (all confusions taken into account)
- Within the plosives, more confusions from voiceless to voiced were found, while no significant effect was found for fricatives
- In medial position, voiced sounds became voiceless more often than voiceless sounds became voiced (chi-square, p < .01) (all confusions taken into account)
- Within the plosives, similar as for initial position, confusions were more often found to be from voiceless to voiced. Unlike the plosives, the fricatives showed more confusions from voiced to voiceless
Observations initial position

- With plosives and fricatives, manner and place are mostly perceived correctly. Most common error is the voicing distinction. /x/ and /h/ difficult.
- For nasals, manner and place are difficult at times. No clear confusion pattern.
- For approximants, manner is very difficult. Confusions are across the board and omissions are common.

Observations medial position

- For plosives and fricatives, manner is mostly perceived correctly. Place causes more problems than in initial position, but is still mainly perceived correctly.
- For approximants, manner and place are problematic.
- For nasals, manner and place cause many problems. There are many confusions with approximants.

Discussion and conclusion: overall

- Even though TE speech intelligibility is remarkably high, considering circumstances, intelligibility is still much lower than for normal laryngeal speech.
- Initial consonants score lowest, final consonants score highest.
- Fricatives score lowest in initial and medial position (also found in American English and Spanish).
- …
Discussion and conclusion: voicing

- Most studies found confusions from voiceless to voiced.
- We found a mix of confusions, both from voiceless to voiced and from voiced to voiceless, with differences between plosives and fricatives and consonant position.
- Explanations might be found in Amsterdam accent and/or anatomy/physiology (hence importance visualization techniques).
- ...

Discussion and conclusions

- Still difficult to give thorough explanation of findings.
- Heterogeneous population with great inter- and intra subject variability.
- Important to find out why some speakers can and other cannot produce certain sounds.
- Hopefully our research will help improve TE speech intelligibility!

The Intervention

Speech therapy for TE speakers?

- At present mainly used for first voice production and prosthesis care.
- But **why** ignore an intelligibility score of +/-75%?
- **Why** settle for moderate speech and not for good to great speech?
- Some small scale studies on TE speech training techniques, but no structured and evidence-based program yet!

Intervention study: two aims

- Does speech therapy based on the specific problems found in TE speakers help improve TE speech intelligibility?
- But also: how do one set up an intervention study and does this set up prove to be the right one?

Subjects

- 9 male Dutch TE speakers.
- Provox voice prosthesis.
- Mean age 64.9 (range 54.11-82.9).
- Mean time after TLE 5.6 (range 0.6-13.7).
In/exclusion criteria

• TLE without reconstruction (inclusion)
• Profound difficulty with producing voice or sound (exclusion)
• Time after TLE minimally 6 months (inclusion)
• Provox 2 (inclusion)
• BUT: beggars can’t be choosers.....

Structure of the program

1. Phoneme level
   • Lesson 1+2: Plosives
   • Lesson 3+4: Fricatives
   • Lesson 5+6: Nasals, /l/, /x/, /h/
2. Discourse level (clear speech)
   • Lesson 7: Explanation
   • Lesson 8: Reading aloud and phrasing
   • Lesson 9: Spontaneous speech

All lessons included auditory feedback

Why like this?

• The phonemes incorporated in the training were found to be most difficult for TE speech in the previous listening experiment
• Transfer from phoneme level to discourse level is necessary if one wishes to improve overall intelligibility
• Clear speech has proven to be beneficial for intelligibility, e.g. in the hard-of-hearing population
• Nine one-hour sessions spread over 5 weeks: Dutch healthcare insurance only pays for 9 sessions

Techniques used

Plosives
1. Auditory discrimination
2. Building up of intra-oral pressure
3. Production (in isolation, and in one- and two-syllable words)

Fricatives
1. Auditory discrimination
2. Sharp articulation, sustaining friction sound longer
3. Production (in isolation, and in one- and two-syllable words)

Nasals
No particular tricks, except using more force:
Keep lips together with more force and for a longer period of time for /m/, pressing tongue more tightly against alveolar ridge for /n/ etc.
Techniques used

/l/
Produce the /l/ with more power and more retroflex

/x/
Start with final /x/ preceded by a vowel. In initial position one can start out from /k/ moving the sound more forward

/i/
Can be learned by producing an inaudible /i/ or a very soft /x/ and prolonging it

Techniques used

Clear speech and phrasing
1. Words are pronounced slower, but not unnaturally slow
2. Speech louder, but no shouting
3. Sounds are not eliminated and articulated well
4. Lively, melodieus, use of stress
5. More frequent breaks, but at the right moments

Pre- and post tests
1. Questionnaires (speakers and relatives)
2. (Phoneme) identification experiment (phonetic listeners)
3. Clinical test (speech pathologists)
4. Rating task (naive listeners)
5. Value judgment (trained listeners)
6. Study specific structured questionnaire

Questionnaire
• Same questionnaire for speaker and close relative before and after training
• Multiple choice question on how they rate their intelligibility in different social situations.
• Speakers also filled out VHI-30 (=Voice Handicap Index, focusing on voice quality) by lack of standardized intelligibility test

(Phoneme) identification task
• 10 phonetically trained listeners
• Speech material consisted of CV, VCV, CVC and SUS
• Online experiment
• Listeners typed in what they perceived

Clinical test
• 10 speech pathologists, all with at least 6 months experience with TE speakers
• Standardized test (dyva) consisting of 25 existing words, including consonant clusters and two-syllable words

<table>
<thead>
<tr>
<th>stimulus</th>
<th>response</th>
<th>targetvocal</th>
<th>targetconson</th>
<th>targetconson</th>
<th>opmerking</th>
</tr>
</thead>
<tbody>
<tr>
<td>plaatsen</td>
<td>pl-</td>
<td>-t-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>zwareboek</td>
<td>z-</td>
<td>-rh-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bruidsme</td>
<td>br-</td>
<td>-dop-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>schoenmer</td>
<td>sch-</td>
<td>-mm-</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(place, thick book, bridal couple, shoe polish)
Rating task

• 10 naïve listeners
• Speech material consisted of reading a story aloud and retelling of the story
• Listeners judged intelligibility and quality according to 17 semantic 7-point scales
• Scales represented the following aspects:
  - Intelligibility, articulation, intonation, tempo, voice, loudness, general qualification

Value judgments

• 4 phonetically trained listeners who also have ample experience with TE speech
• Stimuli consist of voiced and voiceless plosives and fricatives in medial position.
• Task consists of forced choice question and 5 semantic scales, e.g.
  - mooi vs. lelijk (beautiful vs. ugly)
  - krachtig vs. zwak (powerful vs. weak)

Results

Questionnaire: results

<table>
<thead>
<tr>
<th>Category</th>
<th>Pre test</th>
<th>Post test</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects</td>
<td>2.83</td>
<td>3.13</td>
<td>P&lt;.01</td>
</tr>
<tr>
<td>Relatives</td>
<td>3.21</td>
<td>3.35</td>
<td>NS</td>
</tr>
</tbody>
</table>

• On average, subjects scored significantly better in the post test (P<.000)
• All subjects individually scores higher, but in only one person the difference was significant
• For relatives no significant differences could be found between pre and post test
• When comparing subjects and relatives, only in pre test a significant difference can be found, which disappeared in the post test

VHI: average results

<table>
<thead>
<tr>
<th>Category</th>
<th>Pre-test, mean(range)</th>
<th>Post-test, mean(range)</th>
<th>Wilcoxon</th>
</tr>
</thead>
<tbody>
<tr>
<td>VHI-total</td>
<td>19.42(4-57)</td>
<td>16.72(1-46)</td>
<td>NS</td>
</tr>
<tr>
<td>VHI-Functional</td>
<td>13.67(8-22)</td>
<td>11.44(3-17)</td>
<td>NS</td>
</tr>
<tr>
<td>VHI-Emotional</td>
<td>8.44(4-14)</td>
<td>6.33(1-16)</td>
<td>NS</td>
</tr>
<tr>
<td>VHI-Physical</td>
<td>16.78(7-24)</td>
<td>15.67(4-26)</td>
<td>NS</td>
</tr>
</tbody>
</table>

• Both in pre and post test, the E-scores are significantly lower than the P-score (p<.05 for pre-test and p<.01 for post test)
• But no overall significant difference between pre and post test was found

VHI: individual results

• Two subjects show a significant difference between pre and post test
• One shows a significant deterioration
• For subcategories 4 subjects show significant improvement for at least one category, one subject shows a significant deterioration on two categories
**Phoneme identification task: results initial position**

<table>
<thead>
<tr>
<th>Category</th>
<th>Pre test</th>
<th>Post test</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosive</td>
<td>77.1</td>
<td>79.1</td>
<td>NS</td>
</tr>
<tr>
<td>Fricative</td>
<td>49.5</td>
<td>62.8</td>
<td>p &lt; .01</td>
</tr>
<tr>
<td>Nasal</td>
<td>87.7</td>
<td>87.4</td>
<td>NS</td>
</tr>
<tr>
<td>Approximant</td>
<td>77.2</td>
<td>82.8</td>
<td>NS</td>
</tr>
<tr>
<td>Total</td>
<td>67.1</td>
<td>73.7</td>
<td>p &lt; .01</td>
</tr>
</tbody>
</table>

- In initial position, an overall significant improvement is found
- When looking at manners of articulation separately, only fricatives score significantly better

**Phoneme identification task: results medial position**

<table>
<thead>
<tr>
<th>Category</th>
<th>Pre test</th>
<th>Post test</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosive</td>
<td>81.2</td>
<td>86.4</td>
<td>p &lt; .01</td>
</tr>
<tr>
<td>Fricative</td>
<td>62.4</td>
<td>69.3</td>
<td>p &lt; .01</td>
</tr>
<tr>
<td>Nasal</td>
<td>72.8</td>
<td>85.7</td>
<td>p &lt; .01</td>
</tr>
<tr>
<td>Approximant</td>
<td>89.3</td>
<td>92.2</td>
<td>NS</td>
</tr>
<tr>
<td>Total</td>
<td>72.5</td>
<td>79.4</td>
<td>p &lt; .01</td>
</tr>
</tbody>
</table>

- In medial position, an overall significant improvement is found
- When looking at manners of articulation separately, all categories except the approximants show significant improvement

**Phoneme identification task: results voicing**

**Initial**

<table>
<thead>
<tr>
<th>Category</th>
<th>Pre test</th>
<th>Post test</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosive voiceless</td>
<td>80.0</td>
<td>83.0</td>
<td>NS</td>
</tr>
<tr>
<td>Plosive voiced</td>
<td>90.0</td>
<td>94.0</td>
<td>NS</td>
</tr>
<tr>
<td>Fricative voiceless</td>
<td>70.0</td>
<td>77.0</td>
<td>p &lt; .01</td>
</tr>
<tr>
<td>Fricative voiced</td>
<td>85.0</td>
<td>86.0</td>
<td>p &lt; .01</td>
</tr>
<tr>
<td>Total</td>
<td>76.0</td>
<td>81.0</td>
<td>p &lt; .01</td>
</tr>
</tbody>
</table>

- Voicing improves significantly in both initial and medial position, meaning that feature voice was perceived correctly more often
- Only the plosives in initial position do not show a large improvement

**Medial**

<table>
<thead>
<tr>
<th>Category</th>
<th>Pre test</th>
<th>Post test</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plosive voiceless</td>
<td>88.0</td>
<td>92.0</td>
<td>p &lt; .01</td>
</tr>
<tr>
<td>Plosive voiced</td>
<td>89.0</td>
<td>91.0</td>
<td>NS</td>
</tr>
<tr>
<td>Fricative voiceless</td>
<td>68.0</td>
<td>77.0</td>
<td>p &lt; .01</td>
</tr>
<tr>
<td>Fricative voiced</td>
<td>60.0</td>
<td>63.0</td>
<td>NS</td>
</tr>
<tr>
<td>Total</td>
<td>76.0</td>
<td>81.0</td>
<td>p &lt; .01</td>
</tr>
</tbody>
</table>

- In initial position both voiceless and voiced fricatives improve
- In medial position, only the voiceless sounds improve

**Rating task: overall results**

<table>
<thead>
<tr>
<th>Category</th>
<th>Pre test</th>
<th>Post test</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retelling</td>
<td>4.07 (2.97-5.08)</td>
<td>3.83 (2.64-4.29)</td>
<td>p &lt; .01</td>
</tr>
<tr>
<td>Reading aloud</td>
<td>3.91 (2.98-5.07)</td>
<td>3.94 (3.03-4.97)</td>
<td>NS</td>
</tr>
</tbody>
</table>

- Inter-rater reliability pre and post test .886 (Cronbach’s alpha)
- Difference found between the two tasks
- Retelling task scores worse in post test
- Reading aloud task shows slight improvement, but no significant difference can be found

**Rating task: results per category**

<table>
<thead>
<tr>
<th>Category</th>
<th>Pre test</th>
<th>Post test</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Articulation</td>
<td>8.00</td>
<td>7.00</td>
<td>NS</td>
</tr>
<tr>
<td>Emotion</td>
<td>8.00</td>
<td>8.00</td>
<td>NS</td>
</tr>
<tr>
<td>Intonation</td>
<td>8.00</td>
<td>8.00</td>
<td>NS</td>
</tr>
<tr>
<td>Voice</td>
<td>8.00</td>
<td>8.00</td>
<td>NS</td>
</tr>
<tr>
<td>Tempo</td>
<td>8.00</td>
<td>8.00</td>
<td>NS</td>
</tr>
<tr>
<td>General quality</td>
<td>8.00</td>
<td>8.00</td>
<td>NS</td>
</tr>
</tbody>
</table>

- Few significant differences could be found
- Many scales show lower score in post test
- However: intelligibility and articulation either show same score or improvement!
Study specific structured questionnaire: results

- Subjects were handed out evaluation form
- Everyone rated the intervention program quite positively
- All of them claim to have been made aware of their speech, of articulation and what their speech sounds like
- Some claim their speech has become more intelligible

Still to be analyzed

- Data of the clinical test
- Data of the value judgments

Final Discussion

Discussion of results: VHI

- Used by lack of standardized intelligibility test
- Results difficult to explain
- We need to be careful with interpretation: did subjects have intelligibility or voice quality in mind when filling out form?
- Also: too crude an instrument to measure small improvements
- In our opinion, VHI-30 is no suitable test for our purposes

Discussion of results: rating task

- Semantic scales always complicated to investigate
- Still many analyses to do before we understand why post test scores were mostly lower
- However, scores on the intelligibility and articulation scales confirm our findings from the phoneme identification task

Discussion of results: questionnaire

- Although subjects scored higher in the post test, it is difficult with a short questionnaire to obtain significant results
- Subjects were more critical about their own voice than their relatives, but after the intervention, they felt more comfortable with their speech and the difference disappeared
Discussion of results: phoneme identification

- Much more specific test
- At phoneme level significant effects are found for initial and medial position
- In initial position only fricatives score better. Initial position is always harder to hear as it starts abruptly. Approximants show no results, but are difficult to train
- Voicing also improves significantly. Especially fricatives benefit and voiceless plosives and fricatives, which is what intervention focused on most!

Conclusion 1.

- Phoneme training seems to have the desired effect; quite remarkable that after only 5 weeks improvements are found, even in subjects who have been laryngectomized long ago!
- Reading aloud and spontaneous speech probably take longer to improve
- Lower scores on some of the scales might be caused by the fact that subjects are in the middle of a learning process (with speech therapy often things get worse before they get better)

Conclusion 2.

- The fact that subjects were satisfied is important and encouraging
- This program seems suitable to increase intelligibility in laryngectomized individuals and in future most likely will be used for the whole range of TE speakers irrespective of their intelligibility

Recommendations/future

- In view of outcome that VHI-30 is not suitable, there is a clear need for a validated intelligibility test
- Semantic scales are difficult to use for naıve listeners. Another task might have been better
- At 6 months post-intervention, a new evaluation will take place, to allow the assessment of the longer term effects of this rehabilitation program
- Adaptations to the program, if necessary, will be made when on all results are analyzed
- Choices will be made as to what to train and what tests to use to measure results
- In the end, individual results are most important
- It is likely that same results can be obtained in other languages, especially at phoneme level

Questions and Remarks

p.jongmans@nki.nl