Stem-, Spraak- en Taalpathologie
2000, Vol. 9, No.1, pp. 3-21

32.8310/00/0901-003

Betrouwbaarheid en variabiliteit van perceptuele stembeoordelingen door middel van de GRBAS-schaal

M. De Bodt, F.L. Wuyts en P. Van de Heyning

Dienst N.K.O., Hoofd- en Halschirurgie, Revalidatiecentrum voor Communicatiestoornissen, Universitair Ziekenhuis Antwerpen

Perceptuele beoordelingen vormen een essentieel onderdeel van de stembeoordeling ondanks de vele vraagtekens bij de betrouwbaarheid en de variabiliteit van de oordelen. Op basis van verschillende experimenten met de GRBAS-schaal[1] kan een genuanceerde uitspraak worden gedaan over deze aspecten. Test-retest onderzoek toont aan dat de schaal algemeen redelijk betrouwbaar is, maar dat er verschillen zijn volgens de beoordeelde factor. Professionele achtergrond en ervaring hebben slechts een beperkte invloed op de beoordeling. De wijze van aanbieding (auditief of audiovisueel) heeft globaal beschouwd geen invloed op de beoordeling. Het sampletype (vocaal/lopende spraak) daarentegen beïnvloedt de beoordeling wel, in die zin dat vocalen ernstiger worden beoordeeld dan lopende spraak. Deze experimenten tonen aan dat de GRBAS-schaal een bruikbaar instrument is voor de perceptuele beoordeling van stemstoornissen in een klinische multidisciplinaire setting. Anderzijds dient de onderzoeker zich echter ook bewust te zijn van de beperkingen ervan.

Inleiding

De menselijke stem laat zich heel moeilijk vatten in één of meerdere maten of oordelen. Daarom vertrouwen de meeste onderzoekers op een combinatie van metingen en beoordelingen. De indrukwekkende toename aan (technische) onderzoeksmogelijkheden, waaronder met name de akoestische metingen, heeft niet zoveel voordelen opgeleverd voor de dagelijkse klinische praktijk als aanvankelijk werd verwacht (Van de Heyning, Remacle, & Van Cauwenberge, 1996). Bovendien zijn er weinig universeel geaccepteerde en nog minder gestandaardiseerde methodes.

De oudste, frequentst gehanteerde methode is wellicht de perceptuele beoordeling. Het menselijk oor heeft zich als "diagnostisch instrument" weten te handhaven tussen de steeds talrijker wordende instrumentele metingen (De Bodt, Van de Heyning,

Correspondentieadres: M. De Bodt, Universitair Ziekenhuis Antwerpen, dienst N.K.O., Wilrijkstraat 10, 2650 Edegem (België)

Wuyts, & Lambrechts, 1996). Iedereen die met stemstoornissen te maken heeft doet aan één of andere vorm van perceptuele beoordeling, niet in het minst de patiënt zelf. Auditief perceptuele beoordelingen zijn vaak de finale scheidsrechter in klinische beslissingen en zijn de standaard waartegen objectieve metingen worden afgewogen (Kent, 1996). Het naast mekaar bestaan van vele beoordelingssystemen met verschillende terminologie geeft aanleiding tot controversen en kritiek op de methode als dusdanig. Bovendien zijn de meeste perceptuele beoordelingsschalen methodologisch zwak onderbouwd en is overeenstemming tussen beoordelaars eerder zwak.

Kreiman et al. (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993) deden een zeer grondige en kritische studie van de bestaande perceptuele beoordelingsmethodes. Hun literatuurstudie toont aan dat zowel intra- als interbeoordelaarsbetrouwbaarheid sterk fluctueert van studie tot studie. Verschillende factoren dragen daartoe bij: achtergrond en vooroordeel van de luisteraar, de taak zelf, de interactie tussen taak en luisteraar en random error. Op basis van eigen experimenten toonden zij aan dat luisteraars erg variëren in hun oordeel en dat sommige individuele stemmen meer consistent beoordeeld worden dan andere. De variatie is groter bij pathologische dan bij normale stemmen. Opdat perceptuele schalen betekenisvol zouden zijn moeten beoordelaars deze consistent gebruiken. Gerrat et al. (Gerratt, Kreiman, Antonanzas-Baroso, & Berke, 1993) zien goede mogelijkheden in de vervanging van labiele interne standaards door externe standaards (synthetische stimuli bijvoorbeeld). Bij sommige beoordelingssystemen, zoals dat van Hammarberg (1998), werd inmiddels een aanzet gegeven tot het gebruik van een externe standaard.Voor de hier beschreven experimenten werd bewust gebruik gemaakt van de GRBAS-schaal zoals beschreven door Hirano (1981). Deze keuze was vooral ingegeven door de eenvoud van de procedure (minimale training, beperkte afnameduur, beperkte scoringsmogelijkheden), door de internationale erkenning die zij al meer dan vijftien jaar geniet en door de praktische hanteerbaarheid in een multidisciplinaire klinische setting (De Bodt, Van de Heyning, Wuyts, & Lambrechts, 1996).

In dit artikel willen de betrouwbaarheid en de bronnen van variabiliteit van de perceptuele beoordeling (d.m.v. de GRBAS-schaal) onderzoeken. De onderzoeksvragen die wij ons stelden reflecteren concrete aspecten m.b.t. de perceptuele beoordeling. Hoe betrouwbaar zijn de beoordelingen d.m.v. deze schaal? In welke mate beïnvloeden professionele achtergrond en ervaring de perceptuele beoordeling? Welke invloed heeft het sample en de wijze van presentatie van het sample op de beoordeling?

Betrouwbaarheid van de GRBAS schaal

Op de vraag of luisteraars een betrouwbaar oordeel kunnen uitspreken over stemkwaliteit d.m.v. de GRBAS-schaal, komt geen eenduidig antwoord uit de literatuur (Kreiman et al.,1993). Abe et al (Abe, Yonekawa, Ohta, & Imaizumi, 1986) vonden een verschil in reproduceerbaarheid van de beoordeliing van hese stemmen d.m.v. de GRBAS-schaal naargelang van de luisteraars en subschaal. De hoogste reproduceerbaarheid werd gevonden voor "G". Dejonckere et al (Dejonckere, Obbens, De Moor,

& Wieneke, 1993) toonden aan dat de GRBAS-parameters relevant en betrouwbaar zijn door de gunstige combinatie van lage intra- en interbeoordelaarsvariantie en hoge intervoice-variantie.

Wij onderzochten de betouwbaarheid aan de hand van een test-retest experiment (De Bodt, Wuyts, Van de Heyning, & Croux, 1997). Betrouwbaarheid moet hier geïnterpreteerd worden als overeenstemming tussen verschillende beoordelaars op een gegeven tijdstip of door overeenstemming met zichzelf bij het scoren op verschillende momenten (interne consistentie, test/retest). Hiervoor werd de kappa-statistiek aangewend (Altman, 1991; Siegel & Castellan, 1988).

Negen patiënten met uiteenlopende graad van heesheid werden twee maal (met een interval van 2 weken) beoordeeld door een groep van 23 beoordelaars d.m.v. de GRBAS-schaal. Het sample bestond uit een digitale opname van de vocalen /i/ en /a/ en uit een gelezen gestandaardiseerde tekst. De beoordelaarsgroep was samengesteld uit N.K.O.-artsen (n=13) en logopedisten (n=10). De groep werd ook opgedeeld volgens graad van ervaring: ervaren (n=12) en onervaren (n=11) beoordelaars. Tot de groep ervaren beoordelaars werden enkel deze beoordelaars gerekend die meer dan drie jaar ervaring hadden met het beoordelen van stemmen d.m.v. de GRBAS-schaal. Alle beoordelaars kregen identieke instructies en trials m.b.t. het gebruik van de schaal.

Tabel 1 toont de overeenstemming tussen de eerste en tweede beoordeling per groep beoordelaars en voor alle beoordelaars samen.

De kappawaarden die berekend zijn op de mediaanscores kunnen geïnterpreteerd worden aan de hand van tabel 2 (κ=1 betekent perfecte overeenstemming, k=0 betekent overeenstemming puur door het toeval).

De algemene test-retest-betrouwbaarheid is matig (k=0.43). Professionele achtergrond of ervaring beïnvloeden deze betrouwbaarheid niet. De betrouwbaarheid voor "G" is duidelijk beter en dat voor beide professionele groepen. De invloed van ervaring is duidelijker (gaat van 0.50 voor onervaren beoordelaars naar 0.70 voor ervaren luisteraars). Logopedisten geven een betrouwbaarder oordeel over "R" en "B" dan N.K.O-artsen. De betrouwbaarheid voor deze parameters neemt ook toe bij ervaring.

Tabel 1. Test-retest betrouwbaarheid (k) tussen de eerste en tweede score voor ervaren (E) en onervaren (O) beoordelaars, N.K.O-artsen (NKO), logopedisten (LOG) en alle beoordelaars samen (A)

κ	O	E	NKO	LOG	A
G	0.50	0.70	0.62	0.58	0.60
R	0.29	0.40	0.27	0.43	0.35
B	0.29	0.45	0.30	0.46	0.38
A	0.51	0.28	0.42	0.36	0.39
S	0.30	0.36	0.32	0.35	0.34
GRBAS	0.40	0.45	0.41	0.44	0.43

Tabel 2. Interpretatie van kappawaarden naar Altman (Fleiss, 1986)

κ-waarde	Mate van overeenstemming
< .20	"Poor"
0.21 - 0.40	"Fair"
0.41 - 0.60	"Moderate"
0.61 - 0.80	"Good"
0.81 - 1.00	"Very good"

De variabiliteit voor de "A"-parameter neemt echter toe bij meer ervaring, wat erop wijst dat deze parameter vatbaar is voor interpretatiefouten. Globaal genomen zijn ervaren luisteraars het meest betrouwbaar. "G", "R", "B" en "S" zijn vrij betrouwbaar. "A" is dit niet. Deze bevindingen zijn grotendeels conform met die van Dejonckere (Dejonckere et al., 1993; Dejonckere et al., 1996).

De invloed van professionele achtergrond

Hierboven toonden we reeds aan dat professionele achtergrond weinig invloed heeft op de betrouwbaarheid. Anders et al (Anders, Hollien, Hurme, Soninnen, & Wendler, 1988) bestudeerden de effecten van verschillende groepen van luisteraars (foniaters, logopedisten, neus-, keel-, oorartsen en fonetici) op de perceptie van heesheid. Ze vonden kleine maar niet-significante verschillen tussen groepen van luisteraars en besloten dat professionele achtergrond geen majeure rol speelt in de perceptuele beoordeling. De overeenstemming (tabel 3) tussen de verschillende beoordelaars (N.K.O-arts - logopedist) in het hierboven beschreven experiment werd eveneens geanalyseerd aan de hand van de κ-statistiek.

De interbeoordelaarsoveeenstemming is globaal genomen zwak tot matig. Voor alle onderdelen scoren logopedisten iets hoger, wat betekent dat zij iets meer eensgezind oordelen over heesheid dan N.K.O-artsen. Beide groepen werden vergeleken op

Tabel 3. Interbeoordelaarsovereenstemming, verdeeld volgens professionele achtergrond. (NKO= N.K.O.-arts, LOG = logopedist, A= alle beoordelaars samen).

κ	NKO	LOG	A
G	0.45	0.46	0.44
R	0.14	0.25	0.17
B	0.19	0.31	0.21
A	0.26	0.28	0.27
S	0.12	0.12	0.10

Tabel 4. Overeenstemming tussen beoordelaars volgens graad van ervaring: O (onervaren), E (ervaren) en A (alle beoordelaars samen)

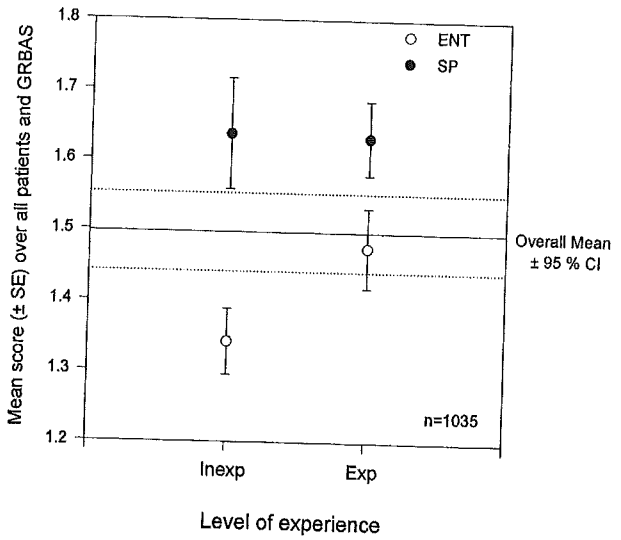
κ	O	E	A
G	0.39	0.49	0.44
R	0.16	0.20	0.17
B	0.21	0.20	0.21
A	0.33	0.27	0.27
S	0.11	0.17	0.10

basis van de mediaanscore d.m.v. de Wilcoxon-toets. Er werd geopteerd voor de mediaan omdat GRBAS-scores gehele getallen zijn en geen tussenliggende waarden kunnen hebben. Significante verschillen tussen beide groepen konden niet worden aangetoond (G:p=0.36, R:p=0.11, B:p=0.18, A:p=0.32 en S:p=0.11).

De invloed van ervaring

Een aantal auteurs beklemtoont het belang van training en ervaring in perceptuele beoordeling (Fex, 1992; Hirano, 1990; Kreiman, Gerratt, & Precoda, 1990; Moller & Starr, 1984; Dejonckere et al., 1996). Moller en Starr (Moller & Starr, 1984) ondervonden dat de variabiliteit in luisteraarsoordelen aanzienlijk kan verminderen door het hanteren van een trainingsprocedure. Gelfer (Gelfer, 1988) vond dat getrainde luisteraars iets meer consistent oordeelden dan ongetrainde. Kreiman et al (Kreiman, Gerratt, & Precoda, 1990) daarentegen toonden aan dat naïeve luisteraars en experten andere perceptuele strategieën vertonen. Experts vertonen minder overeenstemming dan naïeve luisteraars. Bassich & Ludlow (Bassich & Ludlow, 1986) tenslotte rapporteren dat acht uren training nodig zijn om bij onervaren luisteraars 80 % interbeoordelaarsbetrouwbaarheid te bekomen. Dejonckere (Dejonckere et al., 1993) toonde aan dat de GRBAS-schaal een lage interbeoordelaarsvariantie vertoont en dat ervaring met de schaal de interbeoordelaarsovereenstemming significant verbetert (Dejonckere et al., 1996). De intrabeoordelaarscorrelatie is lichtjes beter dan de interbeoordelaarscorrelatie. Wij verdeelden de beoordelaars in het eerder aangehaalde experiment in een ervaren en onervaren groep op basis van het aantal jaren ervaring. Daarna werden deze groepen nog eens opgedeeld volgens discipline (logopedisten, neus-keel-oorartsen). De overeenstemming werd volgens een identieke statistische procedure geanalyseerd.

De resultaten in tabel 4 tonen aan dat ervaren beoordelaars meer overeenstemming vertonen voor G, R en S, maar niet voor B en A. Beide groepen (ervaren en onervaren beoordelaars) werden vergeleken op basis van de mediaanscore d.m.v. de Wilcoxon-toets. Significante verschillen tussen beide groepen konden niet worden aangetoond (G:p=0.18, R:p=0.07, B:p=0.32, A:p=0.32 en S:p=0.06).



Figuur 1. Gemiddelde score voor alle patiënten en GRBAS-parameters volgens niveau van ervaring en professionele achtergrond. ENT=N.K.O.-arts, SP= logopedist, Inexp=onervaren, Exp=ervaren

Tenslotte werd het aspect ervaring ook bekeken voor beide professionele groepen (ervaren logopedisten, ervaren N.K.O.-artsen, onervaren logopedisten, onervaren N.K.O.-artsen). Voor deze groepen werd een gemiddelde berekend gebaseerd op alle toegekende GRBAS-scores aan alle patiënten. Alhoewel deze gemiddelden geen fysiologische betekenis hebben omdat ze over alle stemmen gemiddeld werden, kan men verwachten dat gemiddelden van de vier groepen zich dicht bij het algemeen gemiddelde bevinden. Zoals mag blijken uit figuur 1 was dit niet het geval. Deze figuur toont het algemeen gemiddelde (± 95% confidentie-interval) van alle scores voor alle GRBAS parameters en alle samples van het volledige college van beoordelaaars evenals de subgroep-gemiddelden en hun standaardfouten. Gebaseerd op deze dataset ziet men verschillende tendensen. Over het algemeen scoren ervaren beoordelaars hoger dan onervaren beoordelaars wat erop wijst dat de ernst van de stoornis "overschat" wordt door de ervaren groep en "onderschat" door de onervaren groep. Eenzelfde verschijnsel ziet men bij de vergelijking van logopedisten en N.K.O.-artsen. Hoewel het om kleine verschillen gaat, is het verschil tussen ervaren en onervaren luisteraars kleiner dan dat tussen logopedisten en N.K.O.-artsen, wat suggereert dat professionele achtergrond een grotere impact heeft op perceptuele beoordeling dan ervaring. De factor ervaring is belangrijker bij de N.K.O-groep dan bij de logopedisten . Ervaren N.K.O.-artsen benaderen blijkbaar het dichtst het algemene gemiddelde.

De invloed van wijze van sample-aanbieding: audiovisueel versus auditief

De invloed van de wijze van sample-aanbieding werd, voor zover ons bekend, nog niet eerder bestudeerd. De meeste laboratoriumexperimenten op het vlak van perceptuele beoordeling maken immers exclusief gebruik van auditief aangeboden samples. In een reële onderzoekssituatie beoordeelt men de patiënt echter "live". Videosamples worden nauwelijks gebruikt voor perceptuele beoordelingen van stem. Literatuurstudie (De Bodt, 1997) leert dat in slechts 7 op 563 (=2.7%) publicaties melding wordt gemaakt van videosamples. Verschillende auteurs rapporteren effecten van de wijze van aanbieding op de perceptuele beoordeling van spraakstoornissen (Stephens & Daniloff, 1977; McNutt, Wicki, & Paulsen, 1997; Martin & Haroldson, 1992) Wilson (Wilson, 1987) meent dat voor de meeste aspecten van de communicatie de specifieke luisterconditie weinig verschil uitmaakt met uitzondering van de stemkwaliteit. Wilson baseert zich voor deze uitspraak op een studie van Moller & Starr (Moller & Starr, 1984) die de effecten onderzochten van luistercondities op spraakbeoordeling door getrainde luisteraars bij schisispatiënten. De luistercondities waren resp. "face-to-face", observatie door een spiegel met klankweergave via luidsprekers en beluisteren van een bandopname. Zij vonden geen significante verschillen in de beoordeling van resonantie en articulatie, maar wel voor stembeoordelingen. De stemafwijking wordt als minder ernstig beoordeeld in de live-conditie dan in de twee andere condities (de auditieve en audiovisuele conditie leverde nagenoeg dezelfde oordelen op).

Welk effect de wijze van aanbieding heeft op de perceptuele beoordeling van stem blijft echter onduidelijk. Perceptuele beoordeling d.m.v. de GRBAS-schaal gebeurt op basis van spontane spraak terwijl de onderzoeker de patiënt direct observeert. Het is niet denkbeeldig dat het al dan niet zien van de patiënt de perceptuele beoordeling beïnvloedt. Om dit te onderzoeken, werd een experiment opgezet dat een antwoord moest bieden op de volgende vragen:

(1) Zijn er verschillen tussen zuiver auditieve en audiovisuele beoordelingen van normale en pathologische stemmen aan de hand van de GRBAS-schaal ?
(2) Is er een verschil in beoordeling naargelang van de parameter ?
(3) Is er een verschil in interbeoordelaarsovereenstemming voor beide wijzen van aanbieding?

Methode

34 proefpersonen tussen 8 en 76 jaar met uiteenlopende mate van heesheid werden geselecteerd voor het onderzoek. Van alle proefpersonen werd een hoge kwaliteitsvideo-opname gemaakt die bestond uit volgende onderdelen: datum en naam van de proefpersoon (uitgesproken door de proefpersoon zelf), een gerekte vocaal /a/ gedurende tenminste vijf seconden op comfortabele toonhoogte en luidheid en het lezen van een fonetisch gebalanceerde tekst van 70 woorden. Van de 34 opnames werden 6 (3 mannen, 3 vrouwen) opnames willekeurig geselecteerd als oefenitem. Twintig samples (tabel 5), evenveel mannen als vrouwen, werden geselecteerd door een ervaren luisteraar voor het luisterexperiment. Criteria voor selectie waren de kwaliteit van de klank- en beeldopname en de heterogeniteit aan perceptuele verschillen.

Op basis van de originele band werden één videotape en één audiotape samenge-

Tabel 5. Volgorde van beoordeling bij audiovisuele (AV) en audio (A) aanbieding en bijbehorende patiënt- en stemkarakteristieken (*(M)=man, (V)=vrouw)

Volgorde AV	Volgorde A	Sexe*	Lft	Diagnose	MPT (sec.)	Dynamiek (in dB)	Toonh. bereik (in halve tonen)
1	16	V	46	oedeem van Reinke	27	49	35
2	8	M	42	gezond	17	49	26
3	14	V	43	cyste	10	49	30
4	9	M	75	tumor	12	44	24
5	2	M	57	paralyse in abductie	8	20	14
6	15	V	74	unilat.stemb. parese	–	–	–
7	11	M	8	noduli	–	–	–
8	1	V	29	incomplete sluiting	–	–	–
9	6	M	50	gezond	11	55	30
10	5	V	47	gezond	18	60	36
11	3	V	36	gezond	17	57	42
12	12	M	69	hyperfunctie	13	27	24
13	19	V	38	incomplete sluiting	–	–	–
14	18	M	27	hyperfunctie	–	–	–
15	17	V	32	gezond	17	50	31
16	10	M	11	incomplete sluiting	14	40	19
17	13	V	20	gezond	16	45	30
18	7	M	33	poliep	–	–	–
19	4	V	59	sulcus glottidis	9	24	255
20	20	M	37	paralyse in adductie	7	30	25

steld voor het uiteindelijke experiment. Enkel de gerekte vocaal /a/ en de tekst werden weerhouden om identificatie van de proefpersonen uit te sluiten. De volgorde van de samples voor de audiotape werd gerandomiseerd. Een groep van 20 normaalhorende beoordelaars (laatstejaarsstudenten logopedie) namen vrijwillig deel aan het experi-

ment. Voor het experiment werden 2 sessies georganiseerd, een eerste voor de beoordeling van de audiovisuele samples en een tweede voor de beoordeling van de audiosamples. Bij het begin van de eerste sessie kregen alle beoordelaars nauwkeurige instructies over het GRBAS-beoordelingssysteem en de scoringsprocedure. Om hen vertrouwd te maken met het systeem werden zes samples als oefenitems aangeboden. De twintig samples van het experiment werden in de eerste sessie successief aangeboden aan de groep. De samples werden slechts éénmaal aangeboden. Tussen elk sample werd voldoende tijd geboden om elke beoordelaar te laten scoren. Verbaal contact of discussie tussen beoordelaars werd niet toegestaan. Alle beoordelaars gebruikten hetzelfde scoreblad. Voor de tweede sessie 14 dagen later werd een identieke procedure gevolgd. De beoordelaars werden niet geïnformeerd over de resultaten van de eerste sessie.

Resultaten

Om na te gaan of er verschillen optreden in perceptuele beoordeling tussen audiovisuele en auditieve aanbieding van de samples (vraag 1) werd gebruik gemaakt van de Wilcoxon Matched Pairs Signed Ranks Test (SPSS (Norusis, 1992)). Deze toonde aan dat de audiovisuele scores vrijwel identiek zijn aan de audioscores (p=0.94) wanneer alle scores samen beschouwd worden. De audiovisuele scores zijn identiek aan de audioscores in 1111 oordelen. De audiovisuele score (AV-score) is kleiner (minder ernstig) dan de audioscore (A-score) in 437 oordelen en hoger (ernstiger) in 452 oordelen. Hieruit blijkt dat er geen systematisch verschil bestaat tussen de AV- en de

Tabel 6. Observaties per parameter

Parameter	Rangorde	N	p	Significantie
G	AV < A	72	0.78	NS
	AV = A	248		
	AV > A	80		
R	AV < A	67	0.027	*
	AV = A	239		
	AV > A	94		
B	AV < A	83	0.6	NS
	AV = A	220		
	AV > A	97		
A	AV < A	131	< 0.01	**
	AV = A	214		
	AV > A	55		
S	AV < A	84	0.003	**
	AV = A	126		
	AV > A	190		

NS: niet significant, * 0.05>p>0.01, ** p<0.01
AV < A , AV = A, AV > A (audiovisuele score hoger dan, gelijk aan of kleiner dan audioscore)

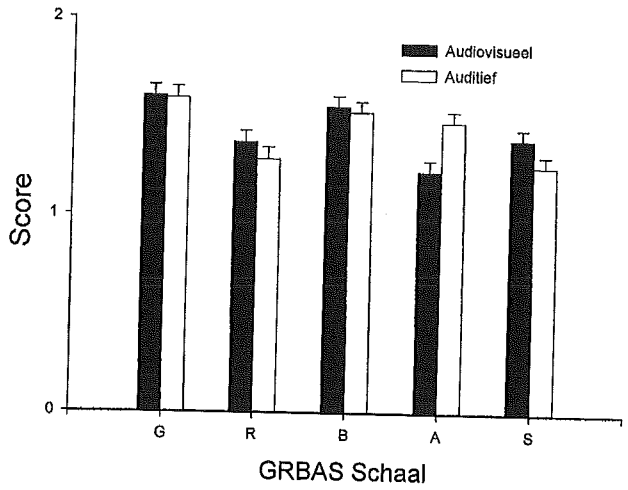
Tabel 7. Interbeoordelaarsovereenstemming (k) voor auditief (AUD) en audiovisueel (AV) aangeboden samples

k	AV	AUD
G	0.44	0.40
R	0.35	0.36
B	0.28	0.25
A	0.20	0.19
S	0.23	0.23

A-scores. Daarnaast werden de verschillende parameters (vraag 2) van de GRBAS-schaal apart geanalyseerd (tabel 6). Voor drie parameters treden significante verschillen op nl. voor "R" (roughness), "A"(asthenic) en "S" (strained). Voor "R" en "S" worden audiovisueel aangeboden samples hoger (dus ernstiger) gescoord dan auditief aangeboden samples. Voor de factor "A" daarentegen worden auditief aangeboden samples hoger (ernstiger) gescoord.

De mate van interbeoordelaarsovereenstemming blijkt uit de kappawaarden (cf. 3.5.4.3) berekend voor beide sessies.

Uit tabel 7 blijkt dat de kappawaarden voor de audiovisuele aanbieding lichtjes hoger zijn dan voor de audio-aanbieding, behalve voor de factor "A". Er blijken statistisch gezien geen significante verschillen te bestaan inzake interbeoordelaarsovereenstemming volgens de wijze van aanbieding van het sample. De overeenstemming is volgens de interpretatie van Altman (1991) behoorlijk (fair) tot matig (moderate) met uitzondering van de "A" (asthenicity) factor bij de audio-aanbieding die als zwak



Figuur 2. Gemiddelde scores volgens wijze van aanbieding

(poor) kan beschouwd worden.

In vergelijking met de studie rond de invloed van ervaring en professionele achtergrond zijn deze kappawaarden lager, hoewel nog steeds betekenisvol. Dit fenomeen kan toegeschreven worden aan het feit dat voor deze studie een beroep werd gedaan op laatstejaarsstudenten logopedie terwijl voor de eerder genoemde studie een beroep werd gedaan op professionelen met een aantal jaren ervaring, wat eens te meer bevestigt dat ervaring een beïnvloedende factor is bij perceptuele beoordeling.

Samenvattend kunnen we stellen dat de wijze waarop het sample ter beoordeling aangeboden wordt (audiovisueel vs. auditief) de perceptuele beoordeling niet beïnvloedt, wanneer alle parameters samen beschouwd worden. Bekijkt men de individuele parameters afzonderlijk (figuur 7), dan vindt men evenmin significante verschillen voor "G" en "B" maar wel voor "R", "S" en "A".

De wijze van sample-aanbieding blijkt dus wel een invloed te hebben op een deel van de parameters. "Roughness" en "Strained" worden ernstiger ingeschat bij audiovisuele aanbieding terwijl "Asthenicity" ernstiger wordt ingeschat bij de auditieve aanbieding. Wat de algemene indruk van de heesheid ("G") en de factor "Breathiness" betreft zijn de verschillen tussen beide aanbiedingsvormen verwaarloosbaar. Het is niet duidelijk wat aan de basis ligt van deze verschillen. Visuele elementen als *houding* en *expressie* versterken misschien de indruk van "R" en "S" terwijl auditieve elementen als een gebrek aan *draagkracht* de indruk versterken van "A". "G" en "B" blijken voor deze elementen eerder ongevoelig te zijn. De betere interbeoordelaarsovereenstemming die hier werd aangetoond voor de audiovisuele aanbieding is analoog aan de bevindingen voor spraakbeoordeling van Stephens & Daniloff (1977) en McNutt et al (1997). Het feit dat geen verschillen geobserveerd worden wanneer de oordelen als één geheel worden beschouwd maar wel wanneer de GRBAS-schaal wordt opgedeeld in zijn individuele parameters, is toe te schrijven aan de inversie van de rangorde van de scores voor de verschillende parameters. Hierdoor worden verschillen in oordeel geneutraliseerd. Deze bevindingen tonen aan dat het vergelijken van experimentele gegevens inzake perceptuele beoordeling met de nodige voorzichtigheid moet gebeuren. Op basis van de aangetoonde betere interbeoordelaarsovereenstemming kan men pleiten voor het algemeen gebruik van de audiovisuele aanbieding.

De invloed van sampletype: lopende spraak versus vocaal

Er bestaat geen eensgezindheid over de vraag op basis van welk sample de perceptuele beoordeling het best kan gebeuren. Hirano (1990) meent dat perceptuele beoordeling het best kan worden gedaan tijdens de anamnese. Wilson (1979) voorziet meerdere taken als: spontaan spreken, lezen, geïsoleerde spraakklanken en automatische taal (bijv. tellen). Askenfeld & Hammarberg (1986) oordelen op basis van ervaring dat lopende spraak meer informatie verschaft over de stemfunctie. Takahashi en Koike (1975) zijn dezelfde mening toegedaan. de Krom (1994) daarentegen besluit op basis van zijn onderzoek naar de betrouwbaarheid van stemkwaliteitsoordelen dat stimuli

op basis van gebonden spraak niet noodzakelijk de voorkeur genieten boven een vocaalstimulus voor de perceptuele beoordeling van "G", "R" en "B". Intra- en interbeoordelaarsovereenstemming worden volgens hem nauwelijks beïnvloed door het stimulus-type. Gegevens over "A" en "S" worden door de Krom (1994) niet gerapporteerd. In een onderzoek bij patiënten met een larynxcarcinoom voor en na radiotherapie stelde Verdonck-de Leeuw (1988) vast dat beoordelaars verschillend oordeelden volgens het type aangeboden sample.

Een tweede element in de discussie rond sampletype is het feit dat akoestische en aërodynamische metingen bijna steeds worden uitgevoerd op basis van vocalen (sustained vowels). Perceptuele beoordeling daarentegen gebeurt meestal op basis van lopende spraak ("running speech"). De resultaten van beide benaderingen worden als complementair beschouwd. Om het effect van samplekeuze op de perceptuele beoordeling na te gaan, werd een experiment opgezet waarin intra- en interbeoordelaarsovereenstemming evenals de ernst van het perceptuele oordeel in functie van het sampletype worden onderzocht.

Methode

Voor dit experiment werden 451 samples (tabel 8) at random geselecteerd uit de "Disordered Voice Database (ver.1.03, october 1994)" die werd samengesteld door de Massachusetts Eye and Ear Infirmery (MEEI) Voice and Speech Lab (Kay Elemetrics Corp. 1994), en verdeeld wordt door Kay Elemetrics Corporation. De database, beschikbaar op CD-ROM, bestaat uit digitale opnames van ongeveer 1400 gezonde en pathologische stemmen, afkomstig van 710 proefpersonen. Een Excell-spreadsheet verschaft alle klinische informatie over deze proefpersonen, o.m. leeftijd, moedertaal, diagnose en akoestische metingen. De CD-ROM bevat een "running speech"- fragment ("Rainbow Passage") (Fairbanks, 1940) en een gerekte vocaal (/a/). De proefpersonen (N=657) lijden aan diverse organische, neurologische, traumatische en psychologische stemstoornissen. Daarnaast bevat de database 53 gezonde stemmen (pathologievrij). De samples werden in vrij veld aangeboden op een comfortabel luidheidsniveau d.m.v. het Kay-CSL-systeem. Vermits het om korte samples gaat, werd voor dit experiment elk sample verschillende keren opnieuw aangeboden tot elke beoordelaar dit voldoende achtte om zijn oordeel te kunnen bepalen.

Drie beoordelaars (twee vrouwen en één man) beoordeelden 902 samples: 451 vocalen en 451 samples van lopende spraak, beide telkens afkomstig van dezelfde proefpersonen. Alle beoordelaars waren normaalhorend. Eén beoordelaar was een ervaren

Tabel 8. Verdeling van de samples over pathologische en gezonde proefpersonen

	mannen	vrouwen	totaal
gezond	21	32	53
pathologische stemmen	165	233	398
totaal	186	265	451

logopedist met meer dan drie jaar ervaring in het gebruik van de GRBAS-schaal. De twee andere beoordelaars waren laatstejaars-logopediestudenten die gedurende 8 uur intensief getraind werden door een ervaren beoordelaar, dit om vertrouwd te worden met de procedure. Voor deze training werd gebuik gemaakt van samples uit dezelfde database die niet werden gebruikt voor het experiment zelf. De beoordeling van de samples aan de hand van het GRBAS-systeem werd gespreid over een aantal sessies zodat niet meer dan 50 samples in één sessie beoordeeld werden. Tussen twee opeenvolgende sessies was er een minimaal interval van 3 dagen. Samples van gezonde en pathologische proefpersonen werden at random gemengd in alle sessies. Lopendespraak-samples en gerekte vocaal-samples werden beoordeeld in aparte sessies. Beoordelaars werden niet geïnformeerd over de klinische gegevens die bij de samples horen.

Resultaten

Om de verschillen in beoordeling na te gaan tussen beide sampletypes werd de interbeoordelaarsovereenstemming bekeken voor elke beoordelaar d.m.v. de k-statistiek (Fleiss, 1986). De k-waarden (tabel 9) verschaffen een idee over de overeenstemming tussen de scores van beide beoordelingssets. Alle k-waarden zijn lager dan .40 wat betekent dat de overeenstemming tussen beoordelingen van lopende-spraak-samples en gerekte-vocaal-samples matig tot gering (fair) is. De test-retestbetrouwbaarheid gaf κ-waarden tussen 0.60 voor "G" en 0.34 voor "S" (De Bodt et al.,1997). Het is duidelijk dat de interne consistentie per beoordelaar relatief klein is bij de beoordeling van het lopende spraak sample en het gerekte vocaal sample van éénzelfde patiënt.

De κ-waarden bij beoordelaar 3 zijn iets hoger dan bij de twee andere beoordelaars wat wellicht kan toegeschreven worden aan het feit dat deze beoordelaar het meest ervaren is in het gebruik van de GRBAS-schaal. Het effect van het sampletype op de beoordeling blijkt consistent te zijn voor beoordelaars 1 en 2 (met vergelijkbare ervaring) vermits zij de parameters "G", "B", "A" en "S" als minder ernstig beoordelen in lopende spraak (LS) dan in een gerekte vocaal (GV). Beoordelaar 3 scoort parameters "R" en "S" hoger (ernstiger) in LS dan in GV (tabel 10).

Een vocaal wordt doorgaans "ernstiger" beoordeeld dan lopende spraak. Deze

Tabel 9. Intrabeoordelaarsovereenstemming tussen *lopende spraak* en *gerekte vocaal* voor elke beoordelaar afzonderlijk (κ)

Parameter	beoordelaar 1	beoordelaar 2	beoordelaar 3
G	0.33	0.36	0.38
R	0.18	0.18	0.22
B	0.14	0.20	0.31
A	0.11	0.11	0.15
S	0.13	0.20	0.11

RS (lopende spraak), SV (vocaal /a/)

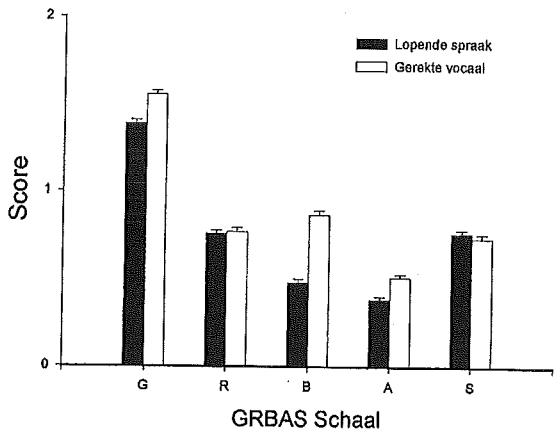
Tabel 10. Observaties per parameter en per beoordelaar

Parameter	Rangorde	Beoord.1	*p*	Beoord. 2	*p*	Beoord. 3	*p*
	GV < LS	56		32		112	
G	GV > LS	151	< .001	172	< .001	88	.062
	GV = LS	244		247		250	
	GV < LS	112		101		134	
R	GV > LS	129	.074	125	.153	98	.023
	GV = LS	210		225		218	
	GV < LS	17		23		91	
B	GV > LS	210	< .001	220	< .001	87	.740
	GV = LS	224		208		272	
	GV < LS	89		48		91	
A	GV > LS	108	.008	154	< .001	93	.071
	GV = LS	254		249		266	
	GV < LS	112		92		163	
S	GV > LS	119	.931	157	< .001	81	< .001
	GV = LS	220		202		206	

LS (lopende spraak), GV (vocaal/a/), GV<LS, GV>LS, GV=LS (scores zijn kleiner, groter of gelijk)

trend wordt geïllustreerd in figuur 3 waar de gemiddelde scores van alle GRBAS-parameters uitgezet zijn voor alle beoordelaars samen en voor alle proefpersonen. Deze gemiddelde scores hebben geen fysiologische betekenis.

Om een idee te krijgen van de interbeoordelaarsovereenstemming vergeleken we de resultaten van alle beoordelaars samen voor LS en GV d.m.v. de k-statistiek. Tabel 11 toont de k-waarden voor LS en GV. Hieruit blijkt dat dat de interbeoordelaarsovereenstemming algemeen beter is bij LS dan in GV voor alle parameters.



Figuur 3. Gemiddelde scores volgens sampletype

Tabel 11. Interbeoordelaarsovereenstemming voor LS en GV

Parameter	Lopende spraak (LS)	Vocaal (GV)
G	0.53	0.45
R	0.38	0.32
B	0.39	0.38
A	0.20	0.20
S	0.30	0.18

Dit experiment stelt twee interessante bevindingen in het licht. (1) Lopende spraak en gerekte vocalen worden duidelijk verschillend beoordeeld. Pathologische condities worden ernstiger ingeschat bij een gerekte vocaal dan bij lopende spraak. Het feit dat de toekenning van ernst meer manifest is bij een vocaal kan verklaard worden door het feit dat de aandacht van de luisteraar bij een gerekte vocaal meer gevestigd is op stem alleen, daar waar in lopende spraak ook andere aspecten van de spraakproductie de aandacht van de beoordelaar weerhouden. Deze bevindingen zijn gedeeltelijk in overeenstemming met deze van Sakata et al (Sakata, Kubuta, Yonekawa, Imaizumi, & Niimi, 1995) die een identieke relatie vonden voor "A" and "S". maar een tegengestelde voor "G" Zij vonden geen significante verschillen voor "R" en "B".

(2) De interbeoordelaarsovereenstemming is beter bij lopende spraak dan bij een gerekte vocaal voor alle GRBAS-parameters. De Krom (1994) onderzocht enkel "G", "R" and en "B" en vond geen effect op intra- of interbeoordelaarsovereenstemming. Zijn studie was gebaseerd op 78 samples. Het is duidelijk dat het sampletype de perceptie van de beoordelaar beinvloedt. Beoordelingen op basis van een gerekte vocaal of lopende spraak kunnen niet zonder meer vergeleken worden met elkaar. De overeenstemming tussen beoordelaars is beter bij lopende spraak dan bij een gerekte vocaal. Het lijkt ons dan ook logisch dat voor de perceptuele beoordeling gebruik gemaakt wordt van lopende spraak, een aanbeveling die we ook terugvonden bij Verdonck-de Leeuw (1998).

Besluit

Tot besluit vatten we de resultaten van de experimenten i.v.m. betrouwbaarheid en de invloed van bronnen van variabiliteit bij perceptuele beoordelingen d.m.v. de GRBAS-schaal nog even samen.

(1) Betrouwbaarheid

Uit het test-retest experiment blijkt dat de betrouwbaarheid algemeen (alle parameters samen) matig (moderate) is, maar aanzienlijk beter (good) is voor de "G" factor. Gelet op het feit dat de k-statistiek rekening houdt met het toeval is dit behoorlijk. De uitge-

breide studie omtrent perceptuele beoordelingsmethoden van Kreiman et al (1993) die pas verscheen nadat deze studie reeds was opgestart, toont aan dat perfecte betrouwbaarheid, zelfs in theorie, zo goed als uitgesloten is. Variabiliteit in luisteraarsoordelen kan volgens deze auteurs gereduceerd worden door de onstabiele interne standaards te vervangen door vaste, externe standaards of referentiestemmen (of "ankerstemmen") voor verschillende stemkwaliteiten. De ontwikkeling van nieuwe schalen moet eveneens zoveel mogelijk contexteffecten elimineren.

(2) Invloed van professionele achtergrond
Perceptuele beoordeling gebeurt door clinici uit verschillende disciplines met uiteenlopende ervaring. Wij hebben kunnen aantonen dat professionele achtergrond (logopedist-N.K.O.-arts) de betrouwbaarheid (test-retest) algemeen weinig beïnvloedt, alhoewel logopedisten iets betrouwbaarder scoren dan N.K.O.-artsen voor de factoren "R" en "B". Er konden geen significante verschillen worden aangetoond tussen beide groepen voor wat de inter-beoordelaarsovereenstemming betreft.

(3) Invloed van ervaring
Ervaren luisteraars beoordelen meer betrouwbaar (consistenter) voor alle parameters behalve voor "A". De interbeoordelaarsovereenstemming is beter bij ervaren beoordelaars voor de factoren "G", "R" en "S" maar niet voor "B" en "A". Wanneer beide groepen vergeleken worden op basis van de mediaanscore, dan kan geen significant verschil worden aangetoond. Wordt ervaring bekeken per professionele groep op basis van de gemiddelden, dan blijken verschillen tussen ervaren en onervaren luisteraars kleiner dan tussen N.K.O.-artsen en logopedisten. Hoewel het om relatief kleine verschillen gaat, blijkt de factor ervaring bij de groep N.K.O.-artsen meer door te wegen dan bij logopedisten.

(4) Invloed van de wijze van aanbieding
Auditief en audiovisueel aangeboden samples worden globaal beschouwd niet anders beoordeeld. Voor individuele parameters treden echter wel significante verschillen op voor "R" (ernstiger beoordeeld bij audio-aanbieding), "S" en "A" (ernstiger beoordeeld bij audiovisuele aanbieding). De interbeoordelaarsovereenstemming is beter bij de audiovisuele aanbieding dan bij de auditieve aanbieding.

(5) Invloed van het sampletype
Het sampletype beïnvloedt het perceptuele oordeel in die zin dat pathologische condities ernstiger worden ingeschat bij een gerekte vocaal dan bij lopende spraak. De interbeoordelaarsovereenstemming is het best bij lopende spraak.

Noten

Perceptuele beoordelingsschaal (Hirano, 1981) met vijf parameters: G (grade), B (breathiness), R (roughness), A (asthenicity) en S (strain) die gescoord worden met

een 4-puntenschaal van 0-3 (normaal, licht gestoord, matig gestoord en ernstig gestoord). De resultaten worden genoteerd d.m.v. een, letter-cijfercombinatie: bijv. $G_2R_2B_0A_0S_1$

Summary

Perceptual evaluation plays an essential role in voice assessment. However, a lot of problems with respect to reliability and variability remain unsolved. Based on a series of experiments with the GRBAS-scale, a number of problems could be addressed. A test-retest experiment showed that the scale is reliable, but also that differences in reliability exist between the subscales. The influence of professional background and level of experience is limited. Sample presentation (audio versus audio-visual presentation) has, in general, no significant effect on the judgements. On the other hand, there is an effect of sample-type (sustained vowel versus running speech): dysphonic voices are judged more severe when judgements are based on sustained vowels. These findings show that the GRBAS-scale is useful for perceptual evaluation in daily clinical settings. On the other hand, clinicians should realize its limitations.

Literatuur

Abe, H., Yonekawa, H., Ohta, F., & Imaizumi, S. (1986). Reproducibilty of Hoarse Voice Psychoacoustic Evaluation. *Japanese Journal of Logopedics and Phoniatrics, 27*, 168-177.

Altman, D.G. (1991). Practical Statistics for Medical Research. London: Chapman & Hall.

Anders, L.C., Hollien, H., Hurme, P., Soninnen, A., & Wendler, J. (1988). Perception of Hoarseness by Several Classes of Listeners. *Folia Phoniatrica, 40*, 91-100.

Askenfeld, A.G. & Hammarberg, B. (1986). Speech Waveform perturbation analysis a perceptual-acoustical comparison of seven measures. *Journal of Speech and Hearing Research, 29*, 50-64.

Bassich, C.J. & Ludlow, C. (1986). The Use of Perceptual Methods by New Clinicians for Assessing Voice Quality. *Journal of Speech and Hearing Disorders, 51*, 125-133.

De Bodt, M.S. (1997). Een onderzoeksmodel voor stemevaluatie. De relatie tussen subjectieve en objectieve parameters in de beoordeling van de normale en pathologische stemfunctie. Doctoraatsproefschrift Universiteit Antwerpen.

De Bodt, M.S., Van de Heyning, P.H., Wuyts, F.L., & Lambrechts, L. (1996). The perceptual evaluation of voice disorders. *Acta Oto-Rhino-Laryngologica Belgica, 50*, 283-291.

De Bodt, M.S., Wuyts, F.L., Van de Heyning, P.H., & Croux, C. (1997). Test-Retest Study of the GRBAS Scale: the Influence of Experience and Professional Background on Perceptual Rating of Voice Quality. *Journal of Voice, 11*, 74-80.

Dde Krom, G. (1994). Consistency and reliability of voice quality ratings for different types of speech fragments. *Journal of Speech and Hearing Research, 37*, 985-1000.

Dejonckere, P.H., Obbens, C., De Moor, G.M., & Wieneke, G.H. (1993). Perceptual Evaluation of Dysphonia: Reliability and Relevance. *Folia Phoniartica, 45*, 76-83.

Dejonckere, P.H., Remacle, M., Fresnel-Elbaz, M., Woisard, V., Crevier-Buchman, L., & Millet, B. (1996). Differential perceptual evaluation of pathological voice quality: reliability

and correlations with acoustic measurements. *Revue Laryngologie, Otologie, Rhinologie, 117*, 219-224.

Fairbanks, G. (1940). *Voice and articulation drillbook.* (2nd ed.) New York: Harper and Brothers.

Fex, S. (1992). Perceptual Evaluation. *Journal of Voice, 6*, 155-158.

Fleiss, J.L. (1986). *The design and analysis of clinical experiments.* New York: John Wiley & Sons.

Gelfer, M.P. (1988). Perceptual Attributes of Voice: Development and Use of Rating Scales. *Journal of Voice, 2*, 320-326.

Gerratt, B.R., Kreiman, J., Antonanzas-Baroso, N., & Berke, G.S. (1993). Comparing internal and External Standards in Voice Quality Judgments. *Journal of Speech and Hearing Research, 36*, 14-20.

Hammarberg, B. (1998). Perception and acoustics of voice disorders – a combined approach. Proceedings of Voicedata 98, Utrecht, 28-29 january 1998.

Hirano, M. (1981). *Clinical Examination of Voice.* Springer Verlag: New York.

Hirano, M. (1990). Clinical Applications of Voice Tests. In NIDCD (Ed.). *Assessment of Speech and Voice Production* (pp. 196-203). Maryland: NIDCD

Kay Elemetrics Corp. (1994). *Disordered Voice Database (ver. 1.03.) of the Massachussetts Eye and Ear Infirmary, Voice and Speech Lab.* Kay Elemetrics 2 Bridgewater Lane, Lincoln Park, NJ 07035-1488.

Kent, R.D. (1996). Hearing and believing: some limits to the auditory-perceptual assessment of speech and voice disorders. *American Jounral of Speech-Language Pathology, 5*, 7-23.

Kreiman, J., Gerratt, B.R., Kempster, G.B., Erman, A., & Berke, G.S. (1993). Perceptual Evaluation of Voice Quality: Review, Tutorial, and a Framework for Future Research. *Journal of Speech and Hearing Research, 36*, 21-40.

Kreiman, J., Gerratt, B.R., & Precoda, K. (1990). Listener Experience and Perception of Voice Quality. *Journal of Speech and Hearing Research, 33*, 103-115.

Martin, R.R. & Haroldson, S.K. (1992). Stuttering ans speech naturalness: audio and audiovisual judgments. *Journal of Speech and Hearing Research, 35*, 512-528.

McNutt, J.C., Wicki, L., & Paulsen, J. (1997). Judgments of phoneme errors under four modes of audio-visual presentation. *Journal of Speech, Language Pathology and Audiology, 15*, 37-42.

Moller, K.T. & Starr, C.D. (1984). The Effects of Listening Conditions on Speech Ratings Obtained in A Clinical Setting. *Cleft Palate Journal, 21(2)*, 65-69.

Norusis, M.J. (1992). *SPSS for Windows. Advanced Statistics. Release 5.* Chicago: SPSS Inc.

Sakata, T., Kubuta, N., Yonekawa, H., Imaizumi, S., & Niimi, S. (1994). GRBAS Evaluation of Running Speech and Sustained Phonations. In Kotby M.N. (Ed.) Proceedings of the XXIII World Congress of the International Association of Logopedics and Phoniatrics. Cairo, 6-10, August 1995.

Siegel, S. & Castellan, J.N. (1988). *Nonparametric statistics for the behavioural sciences.* Mc Graw-Hill.

Stephens, I. & Daniloff, R. (1977). Trouble with /s/: a methodological study of factors affecting the judgement of misarticulated /s/. *Journal of Communication Disorders, 10*, 207-220.

Takahashi, H. & Koike, Y. (1975). Some Perceptual Dimensions and Acoustical Correlates of Pathologic Voices. *Acta Otolaryngolica Supplement, 338*, 1-24.

Van de Heyning, P.H., Remacle, M., & Van Cauwenberge, P. (1996). Functional assessment of voice disorders. Preface. *Acta Oto-Rhino-Laryngologica Belgica, 50*, 25.1

Verdonck-deLeeuw, I. M. (1998). Voice characteristics following radiotherapy: the develop-

ment of a protocol. Academisch Proefschrift. IFOTT, Amsterdam.
Wilson, D.K. (1979). V*oice Problems of Children*. (2nd ed.) Baltimore: Williams & Wilkins.
Wilson, D.K. (1987). *Voice Problems of Children*. (3rd ed.) Baltimore: Williams & Wilkins.