Wim Zonneveld, Hugo Quené and
Willemijn Heeren (eds.)

Sound and Sounds

Studies presented to M.E.H. (Bert) Schouten on the occasion of his 65th birthday

Utrecht Institute of Linguistics OTS

Wim Zonneveld, Hugo Quené and Willemijn Heeren (eds.)

Sound and Sounds

Studies presented to M.E.H. (Bert) Schouten on the occasion of his 65th birthday

This volume contains 17 original contributions by 30 different authors, collected and published to mark the occasion of the 65th birthday of Marten Egbertus Hendrik (Bert) Schouten, teacher and researcher in Phonetics at Utrecht University. A Preface sketches Bert Schouten's many contributions to the field, as an experimental researcher, as a supervisor, as an organizer of highly succesful workshops, and as the chair of the Dutch Association of Phonetic Sciences. The 17 contributions in one way or another all lie in the areas of research in which Bert Schouten has been interested throughout his career: that of the research-enhanced teaching of the pronunciation of English, that of the nature of speech perception, and that of sociophonetics. Contributions are by Gerrit Bloothooft; Elise de Bree and Brigit van der Pas; Desiree Capel, Elise de Bree, Maartje de Klerk, Annemarie Kerkhoff and Frank Wijnen; Johanneke Caspers; Aoju Chen and Vicky Lai; R.P. Clapham, F.J.M. Hilgers, M.W.M. van den Brekel and R.J.J.H. van Son; Janet Grijzenhout; Willemijn Heeren; Vincent J. van Heuven, Nicole N. Broerse and Jos J. A. Pacilly; René Kager and Johannes Schliesser; Hanne Kloots; Brian C. J. Moore; Richard Pastore and Jeremy Gaston; Hugo Quené; Harrie Scholtmeijer; Wang Dandan and Rias van den Doel; and Wim Zonneveld

Utrecht Institute of Linguistics OTS

BSAB-code: 01111946

An exploration into automatic phonological feature evaluation of tracheoesophageal speech

Renee Clapham, Frans Hilgers, Michiel van den Brekel, and Rob van Son

1 Introduction

The move towards automatic measures of speech intelligibility is an attractive response to the drawbacks of perceptual measures of speech intelligibility (e.g. effects of listener familiarity, and time involved in transcribing and scoring). In the clinical situation, assessments of speech intelligibility are often based on perceptual measurements completed by a single clinician. Unlike a human listener, machine evaluation tools are based on acoustic information and little or no language information, much like the ideal listener sought to evaluate speech intelligibility.

Research in the last five years has focused on developing models for predicting global speech intelligibility based on automatic speech recognition tools (see Middag et al., 2008, 2009; Bocklet et al, 2009; Schuster et al., 2006; Van Nuffelen et al. 2009) and studies report strong correlation coefficients between automatic and perceptual measures for large groups of speakers (Van Nuffelen et al., 2009; Windrich et al., 2008). The two tools predominantly reported in the literature are PEAKS (Maier et al., 2009) and the Dutch Intelligibility Assessment (DIA) (Middag et al., 2008, 2009).

This study analyzes the speech material previously presented in Jongmans (Jongmans, 2008) on changes in tracheoesophageal (TE) speech intelligibility as a result of speech therapy. The recordings gathered in Jongmans (2008) contain speech stimuli gathered before and after a 5-week speech therapy program to improve the speech intelligibility of nine TE speakers. In Jongmans (2008), human phoneme recognition of plosives (medial position), fricatives (word initial and medial position) and nasals (medial position) significantly increased after therapy.

We were particularly interested whether the DIA-tool was capable of reflecting these differences. For use in this study the DIA-tool has been adapted to a stand-alone version. The aim of this study is to investigate whether one of the underlying components of the DIA-tool is sensitive to changes in speech as a result of speech therapy.

2 Method

2.1 Speakers

Nine male TE speakers participated in a 5-week speech therapy program to improve speech intelligibility. All speakers had undergone standard total laryngectomy (without additional resection or reconstruction) at least 6 months prior to the therapy program. All speakers used the Provox2 voice prosthesis. This prosthesis is an indwelling device that redirects pulmonary air into the esophageal tract. The same pre-therapy (T0) test stimuli were recorded post-therapy (T1). Due to an error, one speaker read different semantically unpredictable sentences (SUSs) post-therapy in Experiment 2. The data for this speaker were dropped, so data for 8 speakers are used in this experiment.

2.2 Recordings

2.2.1 Experiment 1

Fourteen consonants /p b t d k h x f v s z l n m/ were embedded with the vowels /a i u/ in VCV stimuli. Each consonant was recorded between each of the three vowels (i.e. 14x3 stimuli per speaker). The VCV stimuli were all nonsense syllables. All stimuli were embedded in carrier phrases for the reading task and recordings were made in a sound-treated room. Carrier phases were removed for all recognition tasks.

In the human recognition task the stimuli were presented individually whereas for the machine evaluation all stimuli were concatenated into a single audio file for each speaker (i.e. each speaker had a T0 and T1 audio file).

2.2.2 Experiment 2

Each speaker read a set of 5 SUS before the therapy program (T0) and repeated these sentences after the program (T1) (total 40 different sentence stimuli). To avoid a learning effect for the listeners, each speaker had his own list of SUS. All recordings were made in a sound-treated room.

2.3 Machine recognition

All automatic recognition results were obtained using an adapted version of the online DIA-tool developed by Middag and associates at the University of Ghent (Middag et al., 2008, 2009). The online DIA-tool is designed to predict a speaker's intelligibility based on his/her production of 50 isolated CVC-words. In order to do this, the tool first extracts two speaker dependent feature sets from the recordings, namely phonological features (PLF) and phonemic features (PMF). Using a simple regression model based on human phoneme recognition scores, these features are then converted into an intelligibility score.

For this study the DIA-tool was extended to a stand-alone version and can be run from the command line. Using the same strategy described in Middag et al. (2008, 2009), it can extract speaker feature sets such as the PLFs and PMFs. While the online version is restricted to the 50 isolated words from the a Dutch

intelligibility assessment tool (De Bodt et al., 2006), this offline version can extract the PMFs and PMFs features of any type of speech material as long as a Praat TextGrid-file (Boersma and Weenink, 2009) containing the segmentation and a phonetic lexicon are provided.

With the adapted DIA-tool it is possible to analyze all T0 and T1 speech recordings and derive a PLF and PMF set for every speaker. As described in Middag et al. (2009), there are two kinds of PLFs: positive PLFs and negative PLFs. Positive PLFs represent the extent that a given phonological class supposed to be present for a given phone is supported by the acoustic data during these realizations (i.e. the higher the value the more likely the articulatory aspect was produced how it should have been produced). Negative PLFs represent the degree to which a feature is present when it should not be present (i.e. the lower the value the more likely the articulatory aspect was only produced where it should have been produced). For each feature, recognition values are between 0 and 1 and are derived from posterior probabilities (see Middag et al., 2008, 2009 for more information).

2.3.1 Experiment 1

In Experiment 1 we focus on manner and voicing and by selecting 5 relevant PLFs out of the 26 available. The selected PLFS are: fricative, closure, burst, nasal and voiced. There were two reasons for selecting these features: (1) TE speakers have difficulty with the voice-voiceless distinction for fricative and plosive phonemes, and (2) we can compare the recognition values of the consonant features with the perceptual results reported in Jongmans (2008). We note that as the DIA-tool was designed to predict intelligibility, the PLF set and PMF set have always been used as a whole. Their separate values have not yet been investigated.

2.3.2 Experiment 2

With the adapted version of the DIA-tool, PMF and PLF (negative and positive) information is available for each stimulus. Given the exploratory nature of this study we consider the general behavior of a PLF subset associated with production of the phonemes /p b t d f v s z/. We selected these phonemes as listeners have difficulty identifying alveolar and labial stops and plosives in TE speech intelligibility tasks (Middag et al., 2008). These phonemes are captured in the PLF features fricative, burst, closure (manner), alveolar, labial (place) and voiced.

2.4 Human recognition

2.4.1 Experiment 1

In Jongmans (2008), Jongmans reported the recognition results for these stimuli based on transcriptions by 10 phoneticians (mean age 45;4, range 27-65) with no previous experience with TE speech. The stimuli were presented via an online tool and each listener orthographically transcribed each target consonant for each stimulus. The transcription task included nonsense stimuli for consonant recognition (CV and VCV material), real words for vowel analysis (CVC material) and semantically unpredictable sentences. Stimuli were presented in a block-design. The intra-rater reliability (Cronbach's alpha) for the 10 listeners was 0.796 and 0.767 for

the before and after therapy recordings, respectively (Jongmans, 2008). The results presented in Jongmans (2008) indicate that listeners were able to correctly identify more of the TE speech stimuli recorded after the 5-week speech therapy program than before the program. For consonant recognition, Jongmans reported significant increases (Mc Nemar test) for word initial and word medial fricatives, and word initial plosives and nasals. There was no significant change for approximates. Table 1 displays the results presented in Jongmans (2008) for VCV stimuli (table reproduced with author permission).

Table 1. Recognition results for VCV stimuli presented in Jongmans (2008). The scores represent the mean percentage correct score and ranges.

	Before therapy		After therapy		Sign.
Fricative	62	(49-79)	69	(59-83)	p<.01
Plosive	81	(64-98)	86	(66-97)	p<.01
Nasal	73	(40-92)	86	(72-97)	p<.01
Approximants	90	(57-100)	92	(70-100)	NS
Mean	73	(63-82)	79	(70-86)	p<.01

2.4.2 Experiment 2

Based on the transcriptions by all (10) listeners, we calculated two measures: each speaker's average phoneme error rate (PER) and each speaker's average feature recognition score. The PER for each response is based on identifying phoneme substitutions, deletions or insertions. We calculate the PER as ([substitutions + deletions + insertions]/length). This measure is the phoneme equivalent of the word error rate reported in ASR performance studies and ASR intelligibility studies. In the tables we present these scores as percentages. For analysis of the correlation between PER and DIA-computed intelligibility, we subtracted each speaker's PER from 1 to obtain a phoneme accuracy rate (PAR).

Human feature recognition scores were derived from categorizing correct phoneme recognition according to phonological categories voicing, fricative, plosive, alveolar, and labial for correct recognition of voice, place and manner according to the features. The scores are based on the features in the transcription response versus the feature of the target.

2.5 Statistics

Student t-tests were used where applicable. Non-parametric tests were used when the assumptions of the t-test were not met (e.g., Wilcoxon Signed Ranks test). The level of significance was adjusted using a Bonferroni correction for the number of tests. In Experiment 2 the reported effect sizes are for Cohen's *d*. The strength of the rank order correlation between human recognition and machine recognition was measured using Kendall's Tau. Where the results of parametric and non-parametric tests agree, we present the parametric results within the text. Given the exploratory nature of this study, all comparisons are two-tailed. Analyses were completed with the statistics program R (R Development Core Team, 2010; Gries, 2009).

3 Results: Experiment 1

3.1 Automatic recognition of manner

3.1.1 Positive feature recognition

The mean recognition score for the feature fricative increased by a value of 0.06 after therapy, whereas scores associated with plosives (burst and closure) decreased after therapy (0.03 and 0.06, respectively). A 0.02 decrease in the mean value for recognition of the feature nasal is also seen. On average, the mean sum score for the four features fricative, burst, closure, and nasal for all speakers decreased between T0 and T1 by a value of 0.05. This difference was not significant ($t(8)=0.47$, $p=0.64$).

Although analysis of the mean sum score indicates there is no difference in recognition scores between T0 and T1, given the exploratory nature of this study and the human perception results presented in Jongmans (2008), we completed exploratory pair-wise analysis of features fricative, burst, closure and nasal.

None of the pair-wise comparisons for positive PLFs associated with fricative, plosive and nasal production were significant (fricative: $t(8)=1.65$, $p=0.14$; burst: $t(7)=1.36$, $p=0.22$; closure: $t(7)=2.91$, $p=0.02$; nasal: $t(7)=0.08$, $p=0.93$).

3.1.2 Negative feature recognition

The mean feature recognition values for fricative and plosives increased after therapy. The greatest change was for the feature fricative (0.02 change) whereas the change for burst and closure were smaller (0.02 and 0.01, respectively). The feature not nasal decreased by 0.01.

On average, the mean sum score for the four features decreased by 0.03 for the after speech therapy stimuli. This difference was not significant ($t(8)=1.78$, $p=0.11$). Again, we compared the before and after speech therapy results. These changes were not significant (not fricative: $t(8)=0.76$, $p=0.47$; no burst: $t(8)=1.68$, $p=0.13$; no closure: $t(8)=1.18$, $p=0.27$; not nasal: $t(8)=1.38$, $p=0.10$).

3.2 Automatic recognition of voicing

Due to the lack of a real glottis, TE speakers experience problems with voicing distinctions (eg, Jongmans, 2008). The behavior of the DIA feature voicing was investigated in a separate analysis. This is because the features fricative, burst, closure, and nasal occur only for consonant targets whereas the PLF voiced is not limited to vowels or consonants. For VCV stimuli, the DIA-tool calculates the recognition of voicing over the frames of the two vowel phones and, depending on the target consonant, voiced over the consonant frames. Comparing PLF values for voicing with human recognition values for voicing is not currently possible as this data was not reported in Jongmans (2008).

Although comparing the PLF values for voiced is exploratory, given that the speech training focused on improving production of the voiced-voiceless distinction if speakers were able to accurately control consonant voicing then this may be reflected in the positive and negative PLF voiced values.

Positive recognition (PLF values) for voicing decreased by 0.03 for the T1 stimuli. Pair-wise comparison indicated no significant change in the recognition of the feature voiced between T0 and T1 recordings ($t(8)=1.11$, $p=0.30$). A similar trend was observed for the feature not voiced. The mean value deceased by 0.04 after therapy. Pair-wise comparison of the feature not voiced indicated no significant change in recognition values after therapy ($t(8)=1.08$, $p=0.31$).

4 Results: Experiment 2

4.1 Machine recognition

4.1.1 Positive PLFs

Initial investigation of difference in paired scores of the summed PLF scores for the features fricative, burst, closure, alveolar, labial and voiced, indicated a significant difference between the T0 and T1 scores ($t(7)=4.27$, $p=0.004$, $d=1.39$). The direction of the change does not support a positive change in the group's speech characteristics after therapy.

As seen in Figure 1 and in Table 2, although the mean and mediums have decreased at T1 for the various features, statistical analysis indicates no significant differences between T0 and T1 for voicing ($t(7)= 1.97$, $p=0.09$), closure ($t(7)= 1.99$, $p=0.09$), burst ($t(7)=1.84$, $p=0.11$) and alveolar ($t(7)= 1.80$, $p=0.12$). The comparisons for fricative and labial indicated trends that the pairs differed after speech therapy (fricative: $t(7)=2.63$, $p=0.03$, $d=0.18$; labial: $t(7)=2.47$, $p= 0.04$, $d=0.17$).

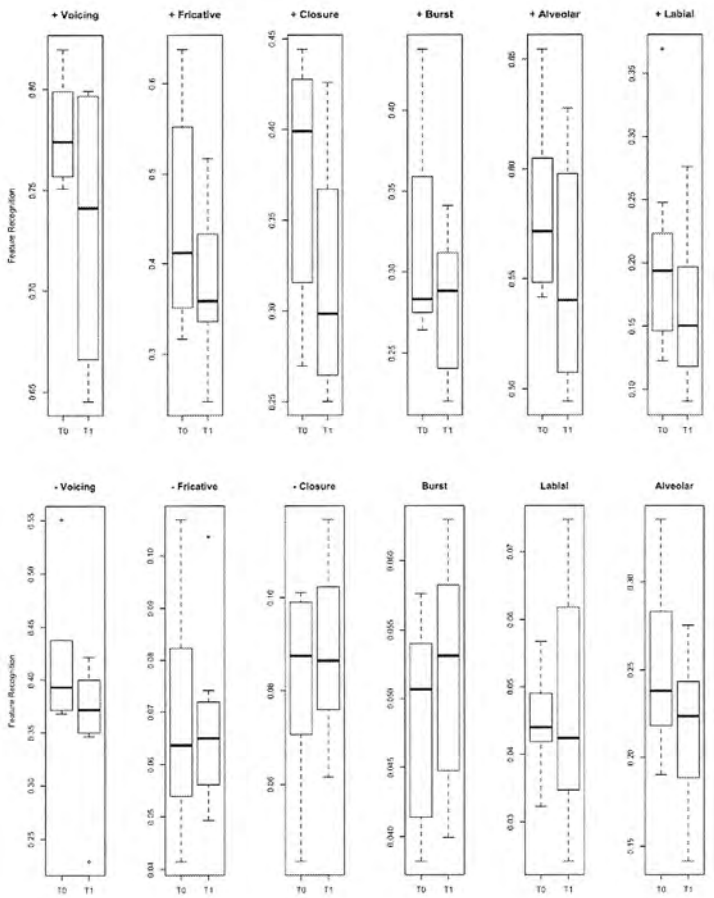

Figure 1. Feature recognition scores before and after therapy. Top: Positive feature recognition. Bottom: Negative feature recognition.

4.1.2 Negative phonological features

Changes in the mean for negative features indicate a mixed pattern: the features not voiced and not alveolar decreased whereas the means of all other features increased (see Table 2). There also appears to be little change in the standard deviation of all the features. Analysis of the difference in paired scores of the summed PLF scores indicated no difference between T0 and T1 ($t(7)=1.48$, $p=0.18$).

We completed pair-wise comparison of the separate features. Statistical analysis indicated no significant differences between T0 and T1 for not fricative ($t(7)=0.26$, $p=0.80$), not burst ($t(7)=1.15$, $p=0.29$), not closure ($t(7)=0.78$, $p=0.46$), not alveolar ($t(7)=1.49$, $p=0.18$), not labial ($t(7)=0.39$, $p=0.71$) and not voiced ($t(7)=2.18$, $p> 0.07$).

4.2 Human recognition

4.2.1 Phoneme error rate

The inter-rater agreement for the 10 listeners was W=0.68 and W=0.60 for T0 and T1, respectively (Kendall's coefficient of concordance (W), $p<0.0001$). The agreement is sufficiently high to use listener averages in subsequent analyzes. On average, phoneme error rates based on sentence transcription were lower post-therapy than pre-therapy. Results of t-test for dependent samples indicates that this difference was significant ($t(7)=3.16$, $p=0.016$, $d = 0.54$).

4.2.2 Human feature recognition

Initial analysis of the summed feature score for human recognition of the features (voicing, fricative, plosive, alveolar, and labial) indicated that the increase in features recognized by the listeners significantly increased from 4.83 (sd: 0.31) to 4.92 (sd: 0.28) ($t(7)=2.97$, $p=0.007$, $d=0.189$). Although the means increased between T0 and T1 (see Table 2), this increase was only significant for the feature alveolar ($t(7)=3.92$, $p=.006$, $d=0.251$). All other features did not reach significance level after the Bonferroni correction (voicing: $t(7)=2.67$, $p=.03$; fricative: $t(7)=2.23$, $p=.06$; burst: $t(7)=3.43$, $p=.011$; labial: $t(7)=0.57$, $p=.59$).

Table 2. Summed speaker feature recognition based on human transcription. The values given are for the mean and standard deviation (in brackets)

Feature	T0	T1
Voiced	0.97 (0.02)	0.99 (0.01)
Fricative	0.92 (0.07)	0.95 (0.07)
Plosive	0.96 (0.03)	0.98 (0.03)
Alveolar	0.06 (0.03)	0.97 (0.03)
Labial	0.88 (0.11)	0.89 (0.12)
Sum Score	4.83 (0.31)	4.92 (0.28)

4.3 Correlation human phoneme recognition with automatic scores

We investigated the strength of the rank-order correlation for three relationships: (1) negative PLFs and human PER, (2) positive PLF and human PER. The correlation was significant for the per speaker summed feature score for negative PLFs (not voiced, not fricative, no closure, no burst, not alveolar, and not labial) and the average PER at both measurement moments (T0: tau = 0.64, $p=.03$; T1 tau = 0.64, $p=.03$).

For the other two comparisons, no significant correlations were found between either the summed positive PLF scores per speaker (no table presented) and the PAR (T0: tau=-0.36, $p=0.27$; T1: tau=0, $p=1$).

5 Discussion

Our preliminary results indicate that PLF values alone may not be able to track changes in speech as measured by listeners in a phoneme recognition task. With a data set of nine speakers and only small differences in feature values before and after therapy, we limit our discussion to trends we observed in the data.

In Jongmans (2008), human recognition of plosives, fricatives and nasals increased after therapy in VCV stimuli. This suggests that production for the phonemes /p b t d k x f v s z n m/ improved as a result of therapy. If this is the case, we would expect that positive PLFs associated with manner of production for fricatives, plosives and nasals would have higher recognition values for the stimuli recorded after speech therapy. This was only the case for the PLF fricative. The features associated with plosives (closure and burst) and the feature nasal decreased. The nature of the transcription task does not allow separation of plosive identification based on the presence of a stop-gap or with plosive release.

Nasal consonants were also under-represented in the speech stimuli compared to fricatives and plosives. While group differences for human recognition of nasals were based on 540 stimuli (2x3 stimuli per speaker, transcribed by 10 listeners per speaker), the PLF value for nasal was based on 54 stimuli (6 stimuli per speaker). This stimuli inequality between means based on all listeners and that for the DIA-tool is not limited to nasals and may partly explain why the PLFs scores did not indicate a difference between the recordings.

As negative PLFs represent the degree that a feature is present outside the target, if speakers became (more) consistent in accurately producing phonetic features at the appropriate moments, we would expect a decrease in negative PLF values after speech therapy. The increase in values for not fricative, no burst and no closure indicates that there was an increase in the articulatory aspects associated with these features outside the phoneme targets. In other words, the speakers had more fricative and plosive characteristics where the targets were not fricatives or plosives. Given that both the positive and negative recognition scores for fricative increased, the clinical implication we can speculate that some speakers generalized fricative production across all of their speech.

We are not able to relate the trends for the feature voiced to listener consonant recognition scores as the PLFs for voiced are calculated across both vowels and consonants. Given the small number of speakers, we cannot say if the decrease in feature recognition for voicing is a difference due to therapy. Based on a confusion matrix for a different group of TE speakers, we know that consonants in medial positions tend to be unvoiced for TE speakers (Jongmans et al., 2006) and the change between T0 and T1 may simply reflect normal TE speaker variation.

A limitation of this study is that although we used an adapted version of the DIA-tool, the underlying neural networks for feature recognition are based on Flemish data. Standard Dutch and Flemish share many speech features, however, there are differences in how some phones are produced, particularly for consonant voicing (Van Compernolle et al., 1991). This means that DIA-based scores may not reflect Dutch listeners' scores. This being said, if the speakers had improved their speech production (which is indicated by the perceptual scores), we would expect systematic changes in the recognition scores. The PLFs, however, may not be robust enough to indicate fine changes in speech production.

The correlation between the sum negative PLF score for the features *not fricative*, *no closure*, *no burst*, *not alveolar*, *not labial*, and *not voiced* and human PER for speaker rank order suggests an association between these two values.

The phoneme error rate scores we used in this study indicate that at the group level, average recognition errors decreased after speech therapy. At a speaker

level, only one speaker appeared to have an increase in error scores after therapy. Jongmans' (2008) results for the same stimuli (scored based on 100% correct sentence transcriptions) also showed that not all speakers' scores increased at T1. At the group level, however, there was a significant increase in sentence transcription scores after therapy.

The DIA-tool's underlying components (the PLFs and PMFs) were not designed for analysis independent of the DIA-tools intelligibility prediction model. By investigating the behavior of the PLFs, we wished to investigate whether feature recognition values could indicate changes in speech production. The trends in our data suggest that although neither the negative or positive PLFs values showed significant improvements in speech characteristics after therapy, negative PLFs associated with production of /p b t d s z f v/ correlated with the human-based phoneme recognition score.

Both human-based scores (PER and feature identification) indicate improved human recognition of TE speech after speakers completed the 5-week speech intelligibility program. This clear result was not replicated using the DIA tool in its standard configuration. This difference may be due to the language of the speech material. The DIA-tool was trained on a different dialect, Flemish, than used by the speakers in Jongmans (2008). Focusing on phonemes that are difficult for listeners to recognize in TE speech (plosives and stops and the voicing distinction) may have biased the results: voicing is more pronounced in Flemish than in Dutch (Windrich et al., 2008).

6 Conclusions

This study was an exploratory investigation of the application of the DIA-tool for the Dutch clinical setting. This study highlights that the DIA-tool intelligibility prediction model and, most likely, for feature recognition components, the feature extraction process should be retrained for speaker dialect differences. We are optimistic that the DIA tool in an adapted form could be used in the clinical setting.

We suggest that future research focus on investigating the clinical application of the DIA-tool. For clinical use an automatic tool must be dialect independent; the feature extraction process should therefore be retrained for a broader group of speakers or dialect-dependent versions should be developed. For clinical use of the tool, a therapist must be able to track changes in the speech intelligibility of a single speaker. Further research into the use of the underlying components to track fine-grained changes in speech characteristics is required.

Acknowledgements

We wish to thank Petra Jongmans for allowing us to reproduce parts of her data. We also gratefully acknowledge the assistance of Catherine Middag for adapting the DIA-tool for use in this study and for her comments on this paper.

References

Bocklet, T., Toy, H., Nöth, E., Schuster, M., Eysholdt, U., Rosanowski, F., Gottwald, F. and Haderlein, T., "Automatic Evaluation of Tracheoesophageal Substitute

Voice: Sustained Vowel versus Standard Text", *Folia Phoniatrica et Logopaedica* 61 (2), 112-116, 2009.

Boersma, P. and Weenink, D., "Praat: doing phonetics by computer (Version 5.1.13)" [Computer program], http://www.praat.org, 2009.

De Bodt, M., Guns, C., and Nuffelen, G.V., "NSVO: Nederlandstalig SpraakVerstaanbaarheidsOnderzoek", *Vlaamse Vereniging voor Logopedisten*, Herentals, Belgium, 2006.

Gries, S.T. *Statistics for linguistics with R*, Berlin: De Gruyter, 2009, pp.217.

Jongmans, P., *The intelligibility of tracheoesophageal speech: an analytic and rehabilitation study*, Ph.D. dissertation. University of Amsterdam, 2008.

Jongmans, P., Hilgers, F. J. M., Pols, L. C. W., & van As-Brooks, C. J., "The intelligibility of tracheoesophageal speech, with an emphasis on the voiced-voiceless distinction", *Logopedics Phoniatrics Vocology* 31, 172-181, 2006.

Maier, A., Haderline, T., Eysholdt, U., Rosanowski, R., Batliner, A., Schuster, M. and Nöth, E. "PEAKS - A system for the automatic evaluation of voice and speech disorders", *Speech Communication* 51, 425-437, 2009.

Middag, M., Martens, J.P., Van Nuffelen, G. and De Bodt, M., "Automated intelligibility assessment of pathological speech using phonological features", *EURASIP*, 2009.

Middag, M., Van Nuffelen, G., Martens, J.P. and De Bodt, M., "Objective intelligibility assessment of pathological speakers", in *Proceedings of Interspeech 2008*, 1745-1748, 2008.

R Development Core Team. "R: A language and environment for statistical computing (Version 2.12.0)" [Computer program], http://www.R-project.org, 2010.

Schuster, M., Haderlein, T., E. Nöth, Lohscheller, J., U. Eysholdt and Rosanowski, F., "Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating", *European Archives Otorhinolaryngology* 263 (2), 188-193, 2006.

Searl, J.P., Carpenter, M.A. and Banta, C.L., "Intelligibility of stops and fricatives in tracheoesophageal speech", *Journal of Communication Disorders* 34, 305-321, 2001.

Van Compernolle, D., Smolders, J., Jaspers, P. and Hellemans, T. "Speaker clustering for dialectic robustness in speaker independent recognition" *Proceedings of Eurospeech 1991*, 723-726, 1991.

Van Nuffelen, G., Middag, C., De Bodt, M., and Martens, J.P., "Speech technology-based assessment of phoneme intelligibility in dysarthria", *International Journal Language and Communication Disorders* 44, 716-730, 2009.

Windrich, M., Maier, A., Kohler, R., Nöth, E., Nkenke, E., Eysholdt, U. and Schuster, M., "Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma", *Folia Phoniatrica et Logopaedica* 60 (3), 151-156, 2008.