Interspeech 2018
2-6 September 2018, Hyderabad

Vowel space as a tool to evaluate articulation problems

Rob J.J.H. van Son[1,2], *Catherine Middag*[3], *Kris Demuynck*[3]

[1]NKI-AVL, Amsterdam; [2]ACLC, University of Amsterdam, The Netherlands
[3]IDLab, University of Ghent, Belgium

r.v.son@nki.nl

Abstract

Treatment for oral tumors can lead to long term changes in the anatomy and physiology of the vocal tract and result in problems with articulation. There are currently no readily available automatic methods to evaluate changes in articulation. We developed a *Praat* script which plots and measures vowel space coverage. The script reproduces speaker specific vowel space use and speaking-style dependent vowel reduction in normal speech from a Dutch corpus. Speaker identity and speaking style explain more than 60% of the variance in the measured area of the vowel triangle. In recordings of patients treated for oral tumors, vowel space use before and after treatment is still significantly correlated. Articulation before and after treatment is evaluated in a listening experiment and from a maximal articulation speed task. Linear models can explain 50-75% of variance in perceptual ratings and relative articulation rate from values at previous recordings and vowel space measures.
Index Terms: pathological speech, vowel space

1. Introduction

After treatment for oral tumors, which involves surgery or radiotherapy, patients often develop problems with speech [1]. In the Netherlands, as in many other countries, patients with head and neck tumors will routinely be seen by speech and language pathologists (SLPs). In the course of therapy, there is a need to quantify and document the quality of speech so both patients and SLPs can evaluate the progress (or not) of the chosen therapy. For voice, there are tools that can give an automatic and objective assessment [2]. However, there is a lack of tools for evaluating articulation and pronunciation beyond perceptual assessments (but see [3]).

In response to questions of SLPs and patients, a project was started to develop tools that might be useful in evaluating articulation and pronunciation. As a starting point, a tool was developed that can visualize and quantify the use of vowel space by speakers based on a recording of (connected) speech. Vowel space parameters are relevant in socio-linguistics [4], speech intelligibility [5], and various pathologies [6, 7, 8, 9, 10]. In contrast with [3], the aim here is to extract only easy to interpret geometric parameters from a short recording. From the vowel realizations in the recording, the effective Vowel Space Area (VSA) is estimated, as are the dimensions of the /a/, /i/, and /u/ corner areas. Figure 1 shows plotted examples from recordings of a patient before and after treatment for oral cancer. The difference between these plots suggests that there must be significant differences in the speech of this patient before and after treatment. However, the tool must be validated before any conclusions can be drawn.

In this paper, first steps are made towards validating the hypothesis that changes in the appearance of the vowel space plots are clinically relevant and reliable, and are useful for patients and SLPs. Three questions are investigated: 1) Do measured vowel space parameters reliably reproduce the known phenomenon of vowel pronunciation? 2) Do these parameters relate to the changes in individual patients? 3) Are vowel space parameters related to clinically relevant aspects of speech?

For 1), vowel reduction as a function of speaker and speaking style is modeled on a corpus of Dutch speech. For 2), it is investigated whether vowel space parameters retain information about speech of patients over time during treatment. For 3), a perceptual evaluation of articulation and measurements of articulation speed in a fast pronunciation task are related to vowel space parameters.

2. Methods

2.1. Vowel Space plots

The *VowelTriangle.praat* script [11] is a freely available *Praat* [12] program which reads or records a section of connected speech and creates a vowel space plot together with some statistics. Speech is searched for likely vowel segments using a method adapted from [13]. Formants are determined with the Split Levinson algorithm [14] in *Praat*. This results in a short (F_1, F_2) trajectory for each detected vowel segment. For each vowel segment, and hence (F_1, F_2) trajectory, the closest approach of the formant trajectories to the positions of the three corner vowels /a/, /i/, and /u/ are determined. All distances are calculated in semitones (d(F'$_i$, F"$_i$) = $12 \cdot Log_2(F'_i/F''_i)$) to normalize between formants and speakers. For male voices, the coordinates of the corners are (Hz) /a/: (850, 1290), /i/: (250, 2100), and /u/: (285, 650), for female voices, /a/: (900, 1435), /i/: (280, 2200), and /u/: (370, 700). These corner points are indicated with crosses in the plots. These points were chosen to enclose the averaged values measured for the isolated vowel samples in the *IFA corpus* [15, 16, 17] for male and female voices. From the corner coordinates, a centroid is determined (geometrical mean of the frequencies). For each vowel segment found, a symbol is plotted at the position of the closest approach to the corner, but only if it lies between the centroid and the corner. It is possible that more than one symbol is plotted for a single vowel segment if it approaches more than one corner close enough. Symbols are colored, green in the /i/ corner, blue in the /u/ corner, and red in the /a/ corner, see Figure 1. For each vowel formant trajectory, the closest approach to the centroid is also plotted in gray. These points are not used in this study.

For quantitative analysis, three axes are defined, one between the centroid and each corner. The positions of the vowels plotted in each corner of the triangle are projected onto the corresponding axis. The mean and standard deviation of the projected positions on these axes are calculated. A length is defined for each axis in vowel space as the distance between the centroid and the mean plus once or twice the standard deviation ($mean + \{1, 2\} \cdot sd$). The triangle spanned by these three axes

10.21437/Interspeech.2018-68

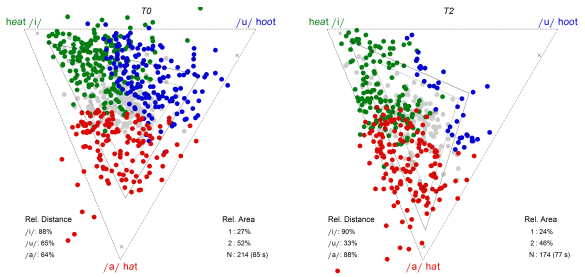

Figure 1: *Example of vowel space plots of a male subject before (left) and one year after (right) treatment for oral cancer. F_1 from top to bottom, F_2 from right to left (in semitones).*

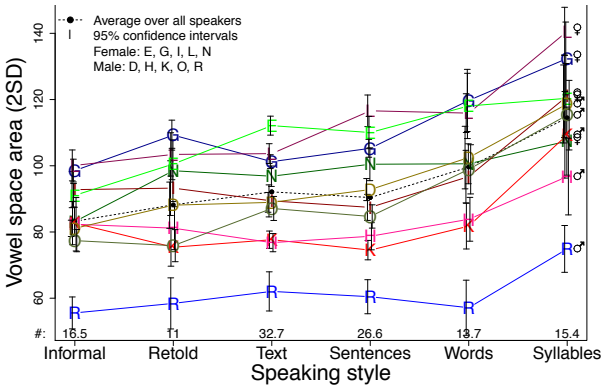

Figure 2: *Average Vowel Space Area (VSA) by speaking style for each speaker. VSA's are calculated on the IFA corpus [16, 17] averaging over the paragraph sized "chunks". Error bars indicate 95% confidence interval (t-test). #: Average # Chunks*

is considered the effective vowel space used. This triangle is drawn in the plot with dotted lines for the $1 \cdot sd$ case and solid lines for the $2 \cdot sd$ case. The VSA of the 1 or 2 sd triangle is a measure for the area of the vowel space as used by the speaker. This is indicated as a percentage of the size of the canonical vowel triangle as given by the corner values. Next to these areas, the plots also contain the relative sizes of the individual axes ($2 \cdot sd$ values). Note that the segmentation method of [13] tends to miss reduced, schwa like, vowels. As a result, the number of detected vowel segments will decline when vowels are pronounced more schwa like. Therefore, the (relative) number of vowel segments too is a measure of average vowel salience. At the bottom right, the number of vowel segments found (N) and the total duration of the recording (s) are also written.

2.2. Speech materials

IFA corpus The *IFA corpus* is used as reference speech from normal speakers [15, 16, 17]. The *IFA corpus* contains recording from 5 male and 5 female native speakers of Dutch, and is freely available (GPL v2 license). For each speaker, recordings are available in different speaking styles: *Informal* speech, a *Retold* story, read aloud *Text*, isolated *Sentences*, multi-syllabic *Words*, isolated *Syllables*, and a few others that are not used in this study. In total, five hours of speech are available.

For each speaker in the *IFA corpus*, the speech was recorded in two sessions on different days (not for *Informal* speech). The material recorded during the two sessions was different on a textual level. The data were both used as is, 1161 *Chunks*, and by concatenating all chunks recorded from the same speaker in the same style in the same session (*Concatenated* set). This set of *Concatenated* recordings contains 100 speech fragments (10 speakers x 5 styles x 2 sessions) which are around ten times as long as the individual chunks.

Patient recordings Existing speech recordings of 30 patients were selected (14 female) who have been treated for oral cancer with surgery and/or radiotherapy (selection was blind to clinical parameters but subjects could read printed text). Each patient has been seen by the SLPs before treatment (T0), 6 months after treatment (T1), and 12 months after treatment (T2). Recordings were made over years and stored as uncompressed audio files. Three different microphones were used (HS5 Samson Headset, Shure SM10A-CN headset with a Blue Icicle USB microphone preamplifier, and Samson Qv10e microphone). Sound quality varies due to sub-optimal recording conditions.

During each visit, a fixed set of recordings was made. Due to technical and administrative causes, recordings from four patients are missing for the first visit after treatment (T1). In total there are 86 recorded sessions available. From each session there is a recording of a short text of neutral content, *80 dappere fietsers* [80 brave cyclists] (∼150 words, 65 seconds) and a word list (36 words, 52 seconds). In addition, at each session the patient was asked to repeat /pataka.../ as fast and as long as possible. This recording allows to evaluate the agility and control of the tongue.

2.3. Listening experiment and plot shape evaluation

Four listeners with experience evaluating speech (SLPs and advanced students) rated a single sentence from each of the 86 text reading recordings (*Ook het weer heeft aan deze tocht meegewerkt*, [The weather has also contributed to this trip]). Listeners were asked to judge the quality of the articulation on a computerized visual analogue scale between *deviant* and *normal*. The order of presentation was randomized separately for each listener. Each listener first heard four practice items which were identical to the last four presented to that listener. The ratings were normalized per listener and averaged over all the listeners.

Vowel triangle plots are meant to be interpreted by humans. To evaluate the concordance between human interpretation and measured aspects of the vowel triangles, 11 naive subjects (unpaid volunteers) evaluated (pseudo-)randomized lists of all 86 plots of vowel triangles from the text readings of the patients. Subjects marked each plot on a scale from 1-10 on being deformed (1) or normal (10), normal being defined as a uniformly filled triangle. The subjects first evaluated a page with 3 labeled plots from normal speakers and 3 labeled plots of strongly deformed vowel triangles illustrative of the most extreme examples in our corpus. Scores of the individual subjects were normalized (mean=0, sd=1) and then averaged over the subjects.

2.4. Statistical analysis

All statistics are done with *R* [18]. Scripts are available in [11]. Linear models are used to get a lower estimate of how much information can be extracted from the VSA measurements. The quality of the models is estimated with the *adjusted* R^2 which measures the fraction of the variance explained by the model, adjusted for the number of factors included. The change in the Akaike Information Criterion (AIC, lower is better) is used as a second measure of relevance [19, 20]. Factors are added pro-

Table 1: *Linear models of Vowel Space Area (VSA) in the IFA corpus. Adjusted R^2 and (AIC) of models predicting the VSA. Sp: Speaker, St: Style, Se: Session. Chunks: Original fragments as present in the IFA corpus, #: 1161, $\overline{\text{VSA}}$: 94±21, mean number of vowel segments $\overline{N}$=67 [6-295]; Concatenated: Chunks concatenated by session, #: 100, $\overline{\text{VSA}}$: 96±20, and $\overline{N}$=683 [120-2234]. $F * G$: F and G and their interactions.*

IFA corpus Model	Chunks	(all $p < 10^{-12}$) Concatenated
Sp	.43 (9707)	.57 (814)
$Sp + St$	.60 (9299)	.80 (743)
$Sp * St$	.64 (9221)	~~.77~~ (~~772~~)
$'' + Se$	.64 (9215)	~~.78~~ (~~770~~)
$'' + Sp * Se$	.66 (9149)	.84 (735)
$'' + St * Se$	.68 (9075)	.92 (661)
$Sp * St * Se$	.69 (9074)	

gressively to a model if they increase both the adjusted R^2 *and* reduce the AIC (if not, ~~*strike-through*~~ is used). At each step, the factor is chosen that increases the adjusted R^2 most. Models that are not statistically significant (p>0.05) are omitted. To validate the generalization of the models, the models were also tested using a Leave-One-Out cross-validation (LOO). When a model generalizes well, the reduction in Mean-Square Error (MSE) due to the model, compared to using the mean, should approach $\widehat{r}^2 = (1 - MSE_{model}/MSE_0) \approx R^2_{adj}$.

3. Results

3.1. *IFA corpus* modeling

The average VSA as a function of speaker and speaking style is plotted in Figure 2. A clear relation between VSA and speaking style is apparent.

Linear models are used to estimate the strength of the relation between the VSA and *Speaker*, Speaking *Style* (excluding *Informal*), and Recording *Session*, see Table 1. The original *Chunks* in the *IFA corpus* are rather small, containing only $\overline{N}$=67 detected vowel segments on average. The concatenated chunks contain on average 10 times as many vowel segments per item ($\overline{N}$=683). Results are presented under column *Concatenated* in Table 1.

The factor *Speaker* alone explains 43% and 57% of the variance (*Chunks* and *Concatenated*, respectively). *Speaker* and *Style* together explain 60% and 80% of the variance, 64% for *Chunks* with the interaction term added too. The remainder of the variability is best explained with a combination of speaking style and session (St*Se). A first likely underlying factor is the difference in size (number of words) between the task in the two sessions. The second likely underlying factor is that the speakers were more familiar with the task in the second session. Increasing the length of the speech fragments used in the analysis, i.e., the number of vowel segments, increases the adjusted R^2 considerably (adding ∼10%).

In the LOO cross-validation test, $\widehat{r}^2$ is 0.65 and 0.78 for *Chunks* and *Concatenated*, respectively. This is close to R^2_{adj} in the *Chunks* case (0.69), but is lower than the 0.92 expected in the *Concatenated* set (see Table 1).

Table 2: *Modeling VSA in Word list task patient recordings (as Table 1). Adjusted R^2 (AIC), using cumulative models in column Model, at T0, T1, and T2. Results for the Word list task only (see text). VSA_t: VSA at time t (0, 1, or 2). Sx: Speaker sex (F or M). *: $p < 0.05$, others: $p < 0.01$*

Mod.	T0	Model	T1	Model	T2
Sx	.64	VSA_0	.27 (210)	VSA_1	.18* (227)
		+VSA_2	.36 (208)	$+Sx$	.31 (223)

3.2. Patient data modeling

3.2.1. Vowel space Area (VSA)

The VSA measurements are somewhat lower and more variable in our patient recordings than in the *Chunks* of the *IFA corpus* ($\overline{\text{VSA}}$: 88±23, cf., Table 1). The average number of vowel segments detected was $\overline{N}$=76 for *Word lists* and $\overline{N}$=203 for *Read Text*. Modeling results of data after treatment were marginal at best for the *Text* reading task and we will focus on the *Word list* task here. Before treatment (T0), the only factor that made a difference was speaker Sex, *Sx*, which behaves as a proxy of speaker identity (results for the *Text* task were comparable at T0). At six months after treatment (T1), the most relevant factor is the vowel space measurement from before treatment (T0). The measurement one year after treatment (T2) has some explanatory power too. At T2, the main factor is the vowel space measure at T1. The next factor of importance is the Speaker sex. None of the interaction terms improve the models.

At both T1 and T2, the largest contribution comes from the vowel space measurement at the preceding recording, T0 or T1. This can be easily understood as the vowel space of the previous recording will capture most of the speaker and task idiosyncrasies of articulation. However, the variance explained is low, 27% at best. This can be increased by adding the vowel space of the T2 for modeling T1, or the speaker sex for modeling T2, explaining a third of the variance in the Word lists. These low values for R^2 at T1 and T2 can possibly be attributed to (uncontrolled) variation in clinical variables in these patients.

In the LOO cross-validation test, the observed $\widehat{r}^2$ are 0.63 for T0, 0.31 for T1, and 0.19 for T2. This is close to the expected value of 0.64 for T0 and 0.36 for T1, but much lower than the expected value of 0.31 for T2 (see Table 2). The model for T2 does not generalize well.

3.2.2. Articulation speed

The articulation rate in the "pataka" task probes the agility of the articulation process. This agility is thought to relate to articulation disorders. However, the articulation rate itself is rather specific (idiosyncratic) for each speaker. To simplify matters, we use the relative articulation rate with respect to the pre-treatment articulation rate (T0). The results were marginal at best for the T1/T0 rate (not shown). The results for the T2/T0 rates were very strong for the text reading task (not shown). The best results were with the shape parameters (*a*, *i*, *u* distances and VSA) pre-treatment (T0). Together these explained 80% of the variance of the relative articulation rate (T2/T0, p<0.001). Adding the i-distance at one year (T2) and the normalized score for plot shape increased this to 88% of the variance for T2/T0. In the LOO test, these models did not generalize ($\widehat{r}^2 \lesssim 0$).

Table 3: *Predicting the Normalized Perceptual Articulation Rating of the listening experiment (as Table 2). Adjusted R^2 (AIC) of models. a_t, i_t, u_t: Measured axis length of the vowel (/a/, /i/, /u/) at time t (0, 1, or 2). Rat_t: Normalized rating at time t. N_0: Vowel segments found at T0. Sx: Sex of speaker (F or M). Ratings were made judging a fragment of the text reading. The highest R^2_{adj} for T1 is 0.70 (47) for $Rat_0 + a_1 * i_1 * u_1$ (not shown). p<0.01 for all models.*

Model Normalized Perceptual Articulation Ratings					
Model	T0	Model	T1	Model	T2
u_0	.47 (49)	Rat_0	.29 (64)	Rat_1	.76 (30)
$+a_0$	.57 (44)	$+i_1$	.42 (59)	$+a_2$	~~.75~~ (~~31~~)
$+Sx$	.67 (36)	$+i_0$	.51 (56)	$+i_2$	~~.77~~ (~~31~~)
$+N_0$	.73 (31)	$+a_1$	.55 (54)	$+u_2$	.79 (30)

3.2.3. Perceptual rating of articulation

The Normalized Articulation Ratings are quite consistent between recordings (see Table 3). 76% of the variance at T2 can be explained by the rating at T1 and 29% of the variance at T1 from the rating before treatment (T0). 67% of the variance in the ratings at T0 can be explained from parameters measured from vowel space and the sex of the speaker. The best linear models for the rating at T1 and T2 both explain around 70% of variance (see Table 3).

A LOO test with the best models showed that $\widehat{r}^2$ values for T0 and T2, 0.67 and 0.79, are close to the expected values 0.73 and 0.79. The $\widehat{r}^2$ value for T1 is worse, 0.36 for an expected 0.70, indicating that the model at T1 does not generalize as well.

3.2.4. Vowel space plot shape evaluation

The vowel space plot shape evaluation *Scores* are modeled using the a, i, and u distances and the VSA, including interactions ($Score_t \sim a_t * u_t * \mathrm{VSA}_t[*i_t]$). Maximal adjusted R^2 (at minimal AIC, all: p<0.05) for T0, T1, and T2 are, respectively, 0.56 (no i_0), 0.58, and 0.66 (no i_2). This shows that the geometrical parameters of the vowel space do describe the visual shape deformation perceptions of the subjects. The LOO test resulted in $\widehat{r}^2$ being somewhat lower than the adjusted R^2, 0.48 and 0.59, for T0 and T2, respectively. For T1, $\widehat{r}^2 \lesssim 0$, i.e., models did not generalize. There was no relation found between the plot shape scores and the results of the listening experiment.

4. Discussion

Figure 1 suggests that measuring the coverage of the vowel triangle might give information about changes in speech (articulation). This would be useful because the *VowelTriangle* script can work on unprocessed recordings. To be useful, it must be shown that the results from the vowel triangle script are consistent, reproduce known features of speech, and are tied to clinically relevant aspects of pathological speech.

An analysis of the *IFA corpus* showed that the relation between vowel reduction, i.e., vowel space coverage, and speaking style, from informal to isolated syllables, are reproduced by the VSA measured by the script (see Figure 2). Depending on the length of the fragments, just the speaker identity and speaking style can explain between 60-80% of the variance in the vowel area coverage (Table 1). This illustrates that the vowel area coverage is highly systematic and associated with speaker identity and speaking style. Adding the recording session boosted the explained variance to 70-90%.

The *VowelTriangle* script could be useful to speech therapists. For that, a visual inspection of the plots should give the relevant impression. A pencil and paper experiment shows that the VSA parameters explain more than half the variance of the normalized scores from naive subjects.

Pathological speech varies more than normal speech and it varies in different ways. Note that the presence of oral tumors can lead to altered speech in patients already before treatment. When measuring changes in vowel articulation, it is important that the link between speech before and after treatment is clear. It was found that characteristics of the vowel triangle after treatment could be modeled by measurements at another moment and by speaker sex explaining a third of the variance of the VSA for reading a list of isolated words (Table 2). The results for the text reading were considerably less consistent. It could be that patients are better able to apply their compensation strategies and preserve their "personal" pronunciation while reading out isolated words than with connected speech.

Finally, to be of clinical use, VSA parameters should be related to clinically relevant characteristics of speech. Evaluation of deviant articulation by experienced listeners is consistent between recordings and, together with vowel space parameters, can explain 60-80% of the variance in ratings. These models are backed by a Leave-One-Out cross-validation. The deviance rating was not correlated to the visual ratings in the pencil and paper experiment.

Changes in a measure related to maximal articulation rate can also be modeled well. Over 80% of the variance in the relative articulation rate one year after treatment (T2) can be explained by the vowel triangle parameters before treatment (T0) and some shape parameters at the time (T2), but only for the running text task. The relative articulation rate 6 months after treatment (T1) cannot be modeled this way. The differences between T1 and T2 found in this study might result from the fact that patients are still recovering 6 months after treatment, which is known from earlier studies [21].

During this study, two technical observations were made. The number of vowels actually detected in speech fragments varied widely, even when the text was identical. Also, differences between male and female speakers were considerable. These observations suggest that vowel detection could be made more robust (e.g., [3, 22]) and the normalization between male and female voices might be improved (e.g., [23, 24]).

5. Conclusions

Vowel space parameters contain relevant information about vowel articulation in normal speakers as well as in the class of speech pathologies found in patients treated for oral cancer. Building predictive models for speech pathologies is outside the scope of this study, but the results obtained with simple linear models suggest that it should be possible to obtain such models using standard machine learning techniques. It is likely that more patient data are needed, that are also better controlled for clinical factors, to construct clinically useful models.

6. Acknowledgements

The Institutional Review Board of the Antonie van Leeuwenhoek Hospital approved the use of speech recordings of patients for this study. The Department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute receives an unrestricted research grant of Atos Medical AB, Hörby, Sweden.

7. References

[1] R. C. Dwivedi, R. A. Kazi, N. Agrawal, C. M. Nutting, P. M. Clarke, C. J. Kerawala, P. H. Rhys-Evans, and K. J. Harrington, "Evaluation of speech outcomes following treatment of oral and oropharyngeal cancers," *Cancer treatment reviews*, vol. 35, no. 5, pp. 417–424, 2009.

[2] B. Barsties and M. De Bodt, "Assessment of voice quality: current state-of-the-art," *Auris Nasus Larynx*, vol. 42, no. 3, pp. 183–188, 2015.

[3] B. H. Story and K. Bunton, "Vowel space density as an indicator of speech performance," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. EL458–EL464, 2017.

[4] E. Jacewicz, R. A. Fox, and J. Salmons, "Vowel space areas across dialects and gender," in *International Congress of Phonetic Sciences*, vol. 16, 2007, pp. 1465–1468.

[5] A. T. Neel, "Vowel space characteristics and vowel identification accuracy," *Journal of Speech, Language, and Hearing Research*, vol. 51, no. 3, pp. 574–585, 2008.

[6] K. Bunton and M. Leddy, "An evaluation of articulatory working space area in vowel production of adults with Down syndrome," *Clinical linguistics & phonetics*, vol. 25, no. 4, pp. 321–334, 2011.

[7] N. Roy, S. L. Nissen, C. Dromey, and S. Sapir, "Articulatory changes in muscle tension dysphonia: evidence of vowel space expansion following manual circumlaryngeal therapy," *Journal of communication disorders*, vol. 42, no. 2, pp. 124–135, 2009.

[8] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech," *Journal of speech, language, and hearing research*, vol. 53, no. 1, pp. 114–125, 2010.

[9] H.-M. Liu, F.-M. Tsao, and P. K. Kuhl, "The effect of reduced vowel working space on speech intelligibility in Mandarin-speaking young adults with cerebral palsy," *The Journal of the Acoustical Society of America*, vol. 117, no. 6, pp. 3879–3889, 2005.

[10] G. S. Turner, K. Tjaden, and G. Weismer, "The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 5, pp. 1001–1013, 1995.

[11] R. J. J. H. van Son, "Vowel Triangle script," https://github.com/robvanson/VowelTriangle, 2018, Also see media files with this proceedings.

[12] P. Boersma and D. Weenink, "Praat: Praat 6.0.36: a system for doing phonetics with the computer," 2017. [Online]. Available: http://www.praat.org

[13] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[14] L. Willems, "Robust formant analysis," *IPO Report*, vol. 529, pp. 1–25, 1986, As implemented in *Praat*.

[15] R. J. J. H. van Son, "The IFA Spoken Language Corpus v1.0," http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/IFAcorpus/, 2001, Accessed: 2017-08-04.

[16] R. J. J. H. van Son, D. Binnenpoorte, H. V. D. Heuvel, and L. C. W. Pols, "The IFA Corpus: a Phonemically Segmented Dutch "Open Source" Speech Database," in *Proceedings of EUROSPEECH 2001 Aalborg*, 2001, pp. 2051–2054. [Online]. Available: http://www.fon.hum.uva.nl/rob/Publications/IFAcorpusEurospeech2001.pdf

[17] R. J. J. H. van Son and L. C. W. Pols, "Structure and access of the open source IFA-Corpus," in *The proceeding of the IRCS Workshop on Linguistic Databases*, 2001, pp. 245–253. [Online]. Available: http://www.fon.hum.uva.nl/rob/Publications/IRCS2001paper.pdf

[18] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: https://www.R-project.org

[19] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, Dec 1974.

[20] K. P. Burnham and D. R. Anderson, "Multimodel inference: understanding aic and bic in model selection," *Sociological methods & research*, vol. 33, no. 2, pp. 261–304, 2004.

[21] L. van der Molen, M. A. van Rossum, C. R. Rasch, L. E. Smeele, and F. J. Hilgers, "Two-year results of a prospective preventive swallowing rehabilitation trial in patients treated with chemoradiation for advanced head and neck cancer," *European Archives of Oto-Rhino-Laryngology*, vol. 271, no. 5, pp. 1257–1270, 2014.

[22] R. P. Clapham, J.-P. Martens, R. J. J. H. van Son, F. J. M. Hilgers, M. M. van den Brekel, and C. Middag, "Computing scores of voice quality and speech intelligibility in tracheoesophageal speech for speech stimuli of varying lengths," *Computer Speech & Language*, vol. 37, pp. 1–10, 2016.

[23] A. C. Lammert and S. S. Narayanan, "On short-time estimation of vocal tract length from formant frequencies," *PloS one*, vol. 10, no. 7, p. e0132193, 2015.

[24] H. Kawahara, T. Kitamura, H. Takemoto, R. Nisimura, and T. Irino, "Vocal tract length estimation based on vowels using a database consisting of 385 speakers and a database with MRI-based vocal tract shape information," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014, pp. 870–874.