UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Computing scores of voice quality and speech intelligibility in tracheoesophageal speech for speech stimuli of varying lengths

Clapham, R.P.; Martens, J.-P.; van Son, R.J.J.H.; Hilgers, F.J.M.; van den Brekel, M.W.M.; Middag, C.

DOI
10.1016/j.csl.2015.10.001

Publication date
2016

Document Version
Final published version

Published in
Computer Speech and Language

License
Article 25fa Dutch Copyright Act

Link to publication

Citation for published version (APA):
Clapham, R. P., Martens, J-P., van Son, R. J. J. H., Hilgers, F. J. M., van den Brekel, M. W. M., & Middag, C. (2016). Computing scores of voice quality and speech intelligibility in tracheoesophageal speech for speech stimuli of varying lengths. *Computer Speech and Language*, *37*, 1-10. https://doi.org/10.1016/j.csl.2015.10.001

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:26 Jun 2022

Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 37 (2016) 1–10

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Computing scores of voice quality and speech intelligibility in tracheoesophageal speech for speech stimuli of varying lengths☆

Renee P. Clapham [a,b,*], Jean-Pierre Martens [c], Rob J.J.H. van Son [b,a], Frans J.M. Hilgers [b,a], Michiel M.W. van den Brekel [b,a], Catherine Middag [c]

[a] *Amsterdam Center for Language and Communication, University of Amsterdam, Spuistraat 210, 1012 VT Amsterdam, The Netherlands*
[b] *Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands*
[c] *Multimedia Lab ELIS, University of Gent, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium*

Received 22 April 2015; received in revised form 19 August 2015; accepted 11 October 2015
Available online 10 November 2015

Abstract

In this paper, automatic assessment models are developed for two perceptual variables: speech intelligibility and voice quality. The models are developed and tested on a corpus of Dutch tracheoesophageal (TE) speakers. In this corpus, each speaker read a text passage of approximately 300 syllables and two speech therapists provided consensus scores for the two perceptual variables. Model accuracy and stability are investigated as a function of the amount of speech that is made available for speaker assessment (clinical setting). Five sets of automatically generated acoustic-phonetic speaker features are employed as model inputs. In Part I, models taking complete feature sets as inputs are compared to models taking only the features which are expected to have sufficient support in the speech available for assessment. In Part II, the impact of phonetic content and stimulus length on the computer-generated scores is investigated. Our general finding is that a text encompassing circa 100 syllables is long enough to achieve close to asymptotic accuracy.
© 2015 Elsevier Ltd. All rights reserved.

Keywords: Laryngectomy; Tracheoesophageal speech; Automatic speech recognition; Speech intelligibility; Voice quality; AMPEX

1. Introduction

The ability to generate automatically computed scores for perceptual variables such as speech intelligibility and voice quality is a relatively recent development in the area of automatic speech and voice evaluation. An advantage of computer-generated scores is that they are not susceptible to extraneous factors, such as listener familiarity with the speaker and differences in listener internal anchors. In the clinical setting, computer-generated scores can be a valuable adjunct to subjective methods of assessment, especially if the evaluation is part of a therapy outcome measurement.

☆ This paper has been recommended for acceptance by Tatsuya Kawahara.

* Corresponding author at: Amsterdam Center for Language and Communication, University of Amsterdam, Spuistraat 210, 1012 VT Amsterdam, The Netherlands. Tel.: +31 0205253805.

E-mail addresses: r.p.clapham@uva.nl (R.P. Clapham), martens@elis.ugent.be (J.-P. Martens), r.v.son@nki.nl (R.J.J.H. van Son), f.hilgers@nki.nl (F.J.M. Hilgers), M.W.M.vandenBrekel@uva.nl (M.M.W. van den Brekel), Catherine.Middag@UGent.be (C. Middag).

http://dx.doi.org/10.1016/j.csl.2015.10.001
0885-2308/© 2015 Elsevier Ltd. All rights reserved.

In fact, prior knowledge of whether a recording is pre-therapy or post-therapy does not influence computed scores as it does with listeners (Ghio et al., 2013) and there is no inter-rater variation for computed scores as there is when perceptual scores are provided by different clinicians.

Computer-generated scores of perceptual variables have predominately been limited to research studies with a focus on developing assessment models, but the methodology is slowly making its way to evaluation studies as a dependent variable (Mayr et al., 2010; Stelzle et al., 2011; Windrich et al., 2008). In most cases, researchers have used speech recordings from existing databases that encompass readings of phonetically balanced texts (e.g. German *Der Nordwind und die Sonne* used in Mayr et al., 2010 and Windrich et al., 2008). In perceptual evaluation of speech intelligibility, some assessments have been developed so that the phonetic material reflects the phoneme frequencies one would expect to measure in long texts from the target language (see review article by Miller, 2013). To our knowledge, the effects of speech stimulus length and phonetic composition on the computed scores has not yet been investigated. There is, however, evidence that improved automatic binary classification (healthy control speakers vs speakers with dysarthria) benefits from more speech material (Bocklet et al., 2013).

The stimulus length varies between research institutes and hospitals as a result of differences in protocol, speaker characteristics (e.g. patient is unable to read the entire text due to reading skills, fatigue or underlying pathology) or both protocol and speaker characteristics. The speech material used across studies within the same institute can also vary and developing distinct assessment models for the various speech materials available is not possible. This motivated us to investigate the impact of phonetic variety and stimulus length on the outputs of automatic assessment models.

The present paper extends our previous work on assessment models for speech and voice quality for speakers treated for head and neck cancer (Clapham et al., 2014; Middag et al., 2014). Where the focus of our previous work was on developing models that perform at a level comparable to that of a human listener when given a sufficiently large amount of speech, the focus of the present work is on developing models that also offer reliable and stable results in a clinical setting where considerably less speech material per subject is available. The main goals are thus (1) to establish strategies for creating more robust models and (2) to offer insight into the minimum amount of speech material needed to attain accurate and stable computer-generated scores with these robust models.

In Section 2 we present the audio stimuli and perceptual evaluation data and describe how the various assessment models were created. We also discuss the methodology used to investigate phonemic variation and model robustness (Part I) and the influence of stimulus length and phonetic composition (Part II). Results from the two experiments are separately listed in the Results section and are discussed as a whole in Section 4.

2. Method

2.1. Audio stimuli

All audio recordings were collected at the Netherlands Cancer Institute (Amsterdam, the Netherlands) as part of previous research studies. As the recordings were gathered over a period of more than 10 years, the recording conditions are partly unknown and most likely differed across studies. Digital audio tape recordings were re-converted into digital form and all recordings were then standardized (sampling frequency of 44.1 kHz, 16-bit linear PCM).

There were recordings of 81 Dutch TE speakers (70 males, 11 females) and all speakers provided informed consent at the time of recording, allowing the recordings to be used for research purposes. Although multiple recordings existed for many speakers, only one recording per speaker (the earliest one) is included in the present study.

All speakers used indwelling voice prostheses (Provox) and read a Dutch text (*80 dappere fietsers*) of neutral content, meaning that the text did not evoke any emotions. The text was divided into six sentences and the average sentence length is 25 words ($SD = 12$, range 13–47) or 47 syllables ($SD = 23$, range 28–88). The text is not phonetically balanced because the recordings stem from research studies which did not require such a balance.

Since we want to study the effects of stimulus length (in syllables), we decided to divide the text into text fragments of almost equal lengths. In a first step we subdivided the longest sentences into two parts by cutting them at a position where a prosodic boundary can be expected. This way we got nine text parts some of which were still too short. In a second step, we therefore merged two short parts into one text fragment. The end result was a set of six text fragments of

Table 1
Illustration of how the nine text parts were recombined into six text fragments of comparable lengths (upper part of table) and of how the six text fragments can be employed to create stimuli of varying lengths (lower part of table).

Text		
Parts:	P1 P2 P3 P4 P5 P6 P7 P8 P9	
Fragments:	F1 (P1); F2 (P2 + P3); F3 (P4 + P8); F4 (P5); F5 (P6); F6 (P7 + P9)	
Combining fragments		
Combination(s):	Single fragment ($n = 6$)	F1; F2; . . .; F6
	2 fragments ($n = 15$)	F1F2; . . .; F5F6
	3 fragments ($n = 20$)	F1F2F3; . . .;
	4 fragments ($n = 15$)	F1F2F3F4; . . .;
	5 fragments ($n = 6$)	F1F2F3F4F5; . . .;

approximately 50 syllables each (mean 47, range 35–54). The upper part of Table 1 illustrates how these text fragments relate to the original sentences. Of the 40 Dutch phonemes, 27 appear in all six fragments.

2.2. Perceptual evaluation

2.2.1. Stimuli

For the auditory–perceptual experiment, we manually extracted the second sentence from the read passage (16 words and 31 syllables) of all speakers as experimental items. We extracted the fifth sentence of 12 randomly chosen speakers as practice items, intended to acquaint the listener with the experimental procedure.

2.2.2. Experimental set-up

Each experimental item was scored by two speech language pathologists, both with extensive clinical experience in the area of TE speech. They individually evaluated the overall voice quality (with descriptors ‘least similar to normal’ and ‘most similar to normal’) and speech intelligibility (descriptors ‘poor’ and ‘good’) on a computerised version of a visual analogue scale in an online self-paced listening experiment. No tick marks were observable during the individual evaluation; the rater moved the cursor along the scale and the final cursor location was saved as a value from 0 to 1000. The latter led to a quantization step that is much smaller than the distinction a listener can make in a statistically confident way, and thus, small enough to justify an interpretation of the discrete value as a continuous score. No speaker information (i.e. gender, age, prosthesis type) was available to the raters.

Scores that differed by more than 125 points between raters were discussed and re-evaluated in a consensus round. In cases where individual scores were within 125 points (corresponds to scores being the same if the scale was converted into a 4-point ordinal scale) the mean score was considered as the consensus score. To aid scoring in the consensus round, major (10% of scale distance) and minor (5% scale distance) tick marks were shown on the scale together with the two individual scores.

2.2.3. Rater agreement and reliability

Before the consensus round, 32 (38%) of the voice quality scores and 46 (54%) of the speech intelligibility scores were within 125 points of each other. The strength of the correlation between the rater’s individual evaluations was significant but low for voice quality (Pearson’s correlation coefficient, PCC = .47, $p < .001$) and adequate for speech intelligibility (PCC = .62, $p < .001$). The intra-class correlation coefficient (for a two-way consistency model) for the variables was fair for voice quality (ICC = .42, $p < .001$) and good for speech intelligibility (ICC = .62, $p < .001$) (Cicchetti, 1994).

2.3. Automatic evaluation tools

Automatic evaluation involves three stages of processing: (1) an acoustic front-end analysis describing the energy and shape of the spectrum of Hamming windowed segments of 25 ms shifted over 10 ms time steps, (2) an analysis of the acoustic information generating global acoustic–phonetic features that characterize the speaker (termed ‘speaker

features') and (3) a prediction of the perceptual variable by means of a regression model. As in our previous work (Middag et al., 2014), we employ an ensemble linear regression model per perceptual variable. Such a model computes the mean of 50 scores generated by 50 small linear regression models. Each small linear model computes the weighted sum of a couple of selected input features and a bias. The features to select and the weights to use are learned on a small randomly selected subset of the training samples. Since an exhaustive search for the best feature subset is computationally prohibitive, the selection strategy works as follows: (a) retain the feature triplets offering the highest model accuracies by means of an exhausive search, (b) extend each retained feature set by adding the feature inducing the largest gain in acccuracy and (c) repeat this until the accuracy of the best feature set saturates or starts to degrade. The model accuracies follow from cross-validation tests on the training subsets.

In the present study, the model inputs are speaker features and these features are organized into five feature sets. We now briefly introduce these feature sets and refer to our previous publications for more details.

2.3.1. Phonological and phonemic features

To derive these features, an automatic speech recognizer matches the acoustic information with the phonetic transcription of the speech via a process of forced speech-to-text alignment. Two types of speaker features can be extracted: phonological features (PLFs) and phonemic features (PMFs). We employ 24 binary phonological properties reflecting manner of articulation (e.g. "burst"), place of articulation (e.g. "bilabial") and voicing (e.g. "voiced"). Each property is either present or absent in the signal, meaning that there are 48 PLFs available to characterize a Dutch speaker: 24 positive and 24 negative features. A high positive PLF indicates that a particular property was present in the intervals it should have been present. A low negative PLF indicates that a particular property was not present in the intervals where it was not supposed to be present.

The PMFs reflect how well phonemes such as /s/, /z/ or /A/ are realized by the speaker. From the likelihoods of the different phonemes in a particular frame one can estimate the posterior probabilities of these phonemes in that frame. The mean of the posterior probabilities of a particular phoneme in the frames aligned with that phoneme is a positive PMF corresponding to that phoneme. A particular PMF thus reflects how well the acoustic properties of a particular phoneme are found in the intervals where that phoneme was uttered. Dutch has 40 phonemes, and therefore, there are 40 PMFs available for characterizing a Dutch speaker.

2.3.2. Alignment-free phonological and alignment-free phonemic features

It is also possible to analyze speech without considering its phonetic transcription. Such an analysis does not involve any speech-to-text alignment and is, therefore, termed 'alignment free'. Alignment-free phonological features (ALF.PLFs) provide information about the phonological properties of the speech signal. As explained in Middag et al. (2010), we use a slightly different phonological feature extractor with 25 instead of 24 phonological outputs in this stage. These phonological outputs are individually analyzed as a function of time (for details see Middag et al., 2010). Per property, this analysis yields 12 features such as "mean value", "mean value of the peaks" and "steepness of the peak onsets". This way, 300 ALF.PLFs (= 25 * 12) are created to characterize the speaker.

In a similar vein, one can also compute ALF.PMFs which provide information about the phonetic properties of the speech signal. Here, a distinction is made between 55 phones (the 40 traditional phonemes plus 6 closures, 6 bursts, 1 glottis and 2 silence symbols) and per phone, analyzing its posterior probability as a function of time now generates six statistical measurements, so that in total 330 ALF.PMFs (6 * 55 = 330) are created to characterize the speaker.

2.3.3. Pitch and voicing related features (AMPEX)

The AMPEX feature extractor generates eight acoustic parameters by means of a built-in auditory model developed by Van Immerseel and Martens (1996). It extracts both voicing-related features (e.g. proportion of voiced frames) and pitch-related features (e.g. jitter). The created features have already been proven successful for the assessment of pathological speech (Moerman et al., 2004, 2015; Clapham et al., 2014). The AMPEX feature extractor is freely available and can be downloaded from the website of the ELIS department at Ghent University.

2.4. Part I: Phonemic variation and model robustness

We compare performances of voice quality and speech intelligibility models that have access to an individual feature sets or combinations of feature sets.

2.4.1. Speech material and sampling procedure

The models are trained and validated using a 5-fold cross validation strategy. A difference with our previous studies (Clapham et al., 2014; Middag et al., 2014) is that we now set aside 1/5 of the stimuli (16 speakers) as test data to be used in Part II of our present study. The remaining stimuli (64 speakers) are divided into five folds: four of which are designated as training data and the remaining one as validation data.

2.4.2. Model inputs

We investigate the performances of full-set models that have access to one or more (up to three) complete speaker feature sets as model inputs and reduced-set models that have only access to preselected features from these feature sets.

As each phonetic feature corresponds to a particular sound (phone or phoneme) and each phonological feature to a set of sounds (both the positive and the negative feature of a sound set is supposed to correspond to the same sound set) it can be characterized by a frequency of occurrence of this sound (set). If a certain sound is not uttered in the analyzed text, the acoustic analysis cannot come up with values for the features corresponding to that sound. In the case of full-set models, we then replace such features by their mean value observed in a big sample of normal speech.

In the case of reduced-set models we only consider features whose expected frequency of occurrence in Dutch (as derived from the phoneme frequencies found in spoken Dutch as reported in Luyckx et al. (2007)) exceeds a threshold of 5%. This meant that (1) we retained only six PMFs (the average posterior probabilities of /@/, /A/, /d/, /n/, /r/ and /t/) and the ALF.PMFs derived from these six PMFs (e.g. “percentage of frames in which /n/ is the phoneme with the highest posterior probability”, “standard deviation of the posterior probability of /n/”), (2) we kept all PLFs except the ten (5 positive and 5 negative) corresponding to the properties ‘approximant’, ‘lateral’, ‘labio-dental’, ‘glottal’ and ‘high’ and (3) we expelled the ALF.PLFs that were derived from the phonological properties ‘nasal vowel’, ‘labio-dental’, ‘glottal’ and ‘palatal’. None of the AMPEX features were expelled as voicing and pitch features are always supported by a sufficient amount of speech.

2.4.3. Performance measures

The primary measure of model performance is the root mean squared error (RMSE). It is defined as the square root of the mean of the squared differences between the predicted (computed) scores and the consensus (perceptual) scores. The goal is to attain a low RMSE. The Pearson Correlation Coefficient (PCC) between the two scores is used as a secondary performance measure, and the goal is to achieve a high PCC.

The Wilcoxon Signed Ranks test is used to establish whether one model significantly outperforms a competitor model. Here it is used to investigate whether the baseline full-set model performance differs significantly from that of the best reduced-set model. A Bonferroni correction for multiple comparisons was used and a conservative p-value ($p < .005$) was deemed statistically significant.

2.5. Part II: Influence of stimulus length and composition

In order to understand why the model scores can be sensitive to the length and the composition of the test stimulus, we recall that many of the speaker features (e.g. components of PMF and PLF) are of the type “average posterior probability of a particular phonological class (either a phone such as /s/ or a phonological class such as ‘plosive’) in the speech intervals realizing a phone of that class”. During model development, we usually employ as much speech material per speaker as possible and means measured on that material are thus bound to approximate the true means for that speaker. During test, however, we want to use short stimuli to reduce the measurement time. In that case, a mean over the test material can significantly deviate from the “true” mean. Moreover, since the relative frequencies of occurrence of the infrequent phonemes of the language strongly depend on the choice of the text, it follows that the number of intervals supporting a particular feature also strongly depends on the text.

In principle, the same reasoning also applies to the alignment-free features, that is, there are two phenomena that can affect the computer-generated scores: variations in the phonetic composition of the text and variations in the length of the text (in phonemes or syllables). If we only consider randomly chosen text fragments, the effects of both phenomena will be strongly correlated as they both give rise to an effect that is expected to be inversely proportional to the square root of the text length. The two experiments we conceived attempt to isolate the two phenomena as much as possible.

Table 2
Performances (PCC reported to two decimal places and RMSE reported to one decimal place) of the best 10 full-set models and the corresponding reduced-set models for voice quality. Also indicated is whether a result is is statistically worse (** $p < 0.005$) than the best result in the column.

Feature sets	Full set		Reduced set	
	RMSE	PCC	RMSE	PCC
PMF	122.7	0.66	126.3	0.61
PMF + AMPEX	122.2	0.66	123.5	0.64
PLF + PMF	125.2	0.64	130.0	0.58
PMF + ALF.PLF	127.6	0.63	124.8	0.64
PMF + ALF.PMF	129.1	0.58	130.8	0.58
PLF + ALF.PLF	138.9**	0.53	137.3**	0.54
PMF + AMPEX + ALF.PLF	122.2	0.66	124.1	0.65
PMF + AMPEX + PLF	124.5	0.64	128.9	0.59
PMF + AMPEX + ALF.PMF	127.3	0.63	129.9	0.58
PLF + ALF.PLF + AMPEX	138.9**	0.53	136.8**	0.54

2.5.1. Models and speech material

We predict consensus scores of voice quality and speech intelligibility with the best-performing models identified in Part I. These models were developed using the readings of complete paragraphs, but are here used to compute scores from speech samples of different lengths.

The test material in this Part is the material that was set aside in Part I. As stated in Section 2.1, the paragraph was divided into six text fragments (F1–F6) of approximately 50 syllables each. Per speaker, we compute scores for all possible stimuli we can construct by combining one up to five text fragments: 6, 15, 20, 15 and 6 stimuli containing 1, 2, 3, 4, and 5 text fragments respectively. The number of fragments in a stimulus is referred to as the stimulus length (in text fragments). Table 1 lists the individual fragments and examples of fragmet combinations of different lengths.

2.5.2. Score processing

Per speaker and per stimulus length, we measure the standard deviation (*SD*) of the scores of all stimuli of that length. The SDs reveal the effect of the phonetic composition on the computed scores for a stimulus of that length. Obviously, we would not have been able to estimate the SD for a stimulus length of six as there is only one stimulus consisting of all six text fragments. This explains why this stimulus length was not included in the experiment. It is important to note here that stimuli of length 2 or larger are not independent of each other as they always share at least one text fragment with another stimulus. However, in spite of this, the measured SDs are bound to provide useful information on the effect of phonetic composition as a function of stimulus length.

In order to assess the effect of the length under the assumption that phonetic variation could be eliminated per length (e.g., by carefully choosing texts for which the phoneme frequencies are identical), we consider the mean of the scores of all stimuli of a particular length provided by a particular speaker as the computer generated speaker score. Using these scores we then compute the RMSE and PCC for that length.

3. Results

3.1. Part I: Phonemic variation and model robustness

3.1.1. Voice quality

Table 2 lists the performances of the 10 best full-set models and the corresponding reduced-set models when applied to full paragraphs. The full-set models PMF + AMPEX (RMSE = 122.2) and PMF + AMPEX + ALF.PLF (RMSE = 122.2) attain the highest accuracy, but the differences across models are small: only 2 of the 10 models of the same type (full-set/reduced-set) perform significantly worse (Wilcoxon, $p < .005$) than the best model of that type. There are no statistically significant differences between the full-set and the reduced-set models of a particular combination of feature sets.

Observe that a combination of three feature sets (e.g. PMF + AMPEX + PLF) does not necessarily lead to a better model than a combination of only two of these feature sets (e.g. PMF + AMPEX), despite the fact that the model found

Table 3
Performances (PCC reported to two decimal places and RMSE reported to one decimal place) of the best 10 full-set models and the corresponding reduced-set models for speech intelligibility. Also indicated is whether a result is statistically worse (**$p < 0.005$) than the best result in the column.

Model features	Full set		Reduced set	
	RMSE	PCC	RMSE	PCC
PMF	98.3	0.67	98.6	0.66
PLF + PMF	97.8	0.67	100.9	0.64
PMF + AMPEX	98.8	0.67	100.0	0.67
PMF + ALF.PLF	100.9	0.66	114.4	0.62
PMF + ALF.PMF	101.3	0.65	102.4	0.63
PLF + AMPEX	117.3**	0.44	116.1**	0.45
PLF + ALF.PMF	117.3**	0.45	123.6**	0.40
PMF + AMPEX + PLF	97.4	0.67	100.2	0.65
PMF + AMPEX + ALF.PLF	98.8	0.67	110.7	0.63
PMF + AMPEX + ALF.PMF	100.8	0.66	102.5	0.63

Table 4
Mean SD and SD-range per perceptual variable of the speaker scores per combination of text fragments.

Fragment combination	*n*	Voice quality		Speech intelligibility	
		Mean	Range	Mean	Range
Single fragment	6	62	16–176	56	18–145
2 fragments	15	39	10–182	37	14–104
3 fragments	20	30	7–225	27	9–90
4 fragments	15	23	5–261	20	8–73
5 fragments	6	15	2–198	12	3–41

in the latter case is an example of an eligible model that could have been found in the first case. This observation reveals the sub-optimality of the feature selection process incorporated in the regression model training. The degree of sub-optimality is expected to increase with the number of features.

3.1.2. Speech intelligibility

For speech intelligibility, the full-set PMF + AMPEX + PLF model is the strongest performing model (RMSE = 97.4) but the strongest performing reduced-set PMF model (RMSE = 98.6) is not far behind (see Table 3). Nevertheless, since the PMF + AMPEX model comes very close to the best model in the two conditions and since it was also the chosen model for voice quality, we will consider PMF + AMPEX as the baseline feature set combination in Part II. As seen for voice quality, only 2 of the 10 competitors of the same type perform statistically worse than the baseline and there are no statistically significant differences between the reduced-set and the full-set models for a given feature set combination.

3.2. Part II: Effects of stimulus composition and length

3.2.1. Influence of phonetic composition

To investigate the influence of phonetic composition on the computed scores, we considered the reduced-set PMF + AMPEX model. Per perceptual variable, per speaker and per stimulus length, we record the SD of the scores across stimuli. The statistics of these SDs across speakers are listed in Table 4. The mean SD clearly decreases with an increasing stimulus length, but the SD-range only seems to decrease for speech intelligibility and not for voice quality.

3.2.2. Influence of stimulus length

To isolate the influence of stimulus length on the reliability of the speaker features, we consider, per variable, speaker and stimulus length, the mean score found across stimuli. Fig. 1 shows the RMSE and PCC obtained by comparing these means to the consensus scores. For both perceptual variables, the accuracy improves (RMSE decreases and PCC

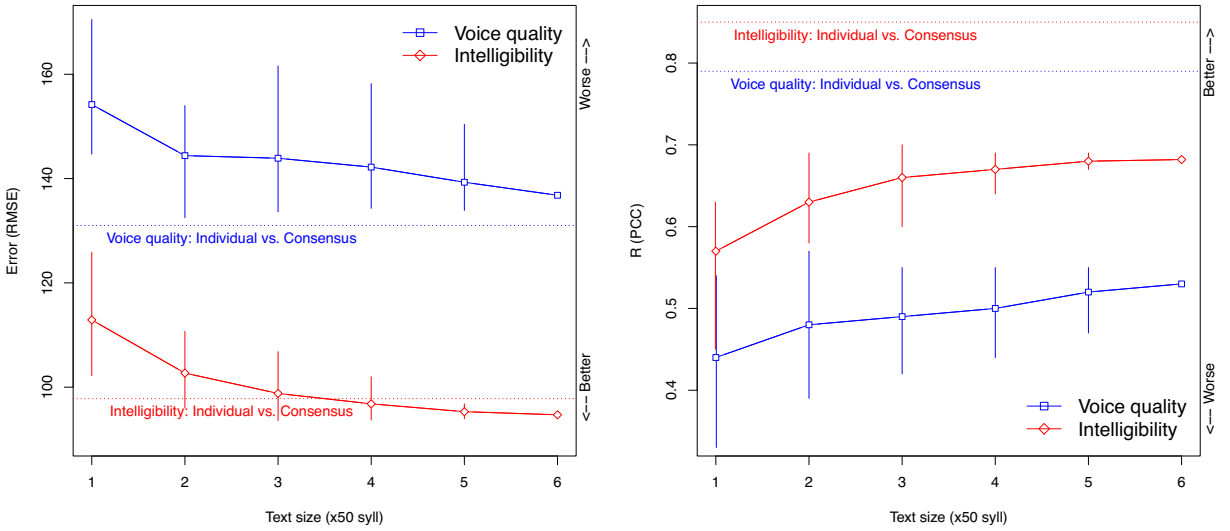

Fig. 1. Accuracy of the mean score (squares/diamonds) and range (verticle lines) across groups (RMSE and PCC) for voice quality and speech intelligibility as a function of the number of fragment combinations (= text size). For comparison, we include the average RMSE and PCC values for the individual SLPs versus the consensus scores (horizontal lines).

increases) when the test stimulus is longer. The improvement is significant and close to 10% when going from 47 syllables (1 text fragment) to 94 syllables (2 text fragments). The improvement caused by adding a third text fragment is not statistically significant anymore.

We also inspected the speaker scores generated for single text fragments. We found that for both perceptual variables, fragment F4 consistently gave rise to low RMSE and high PCC values (RMSE = 144.7 for voicing and RMSE = 106.19 for speech intelligibility). When two fragments are considered, the best combination is F2F4 for both variables (RMSE = 134 for voice quality and RMSE = 96 for speech intelligibility). It happens that F4 comprises a high number of distinct phonemes and syllables and a number of shorter phrases (Dutch: "... en hielp gedurende vijf dagen mee bij het plakken van banden, het maken van gebroken kettingen, het verzorgen van slaapgelegenheid en het opsporen van verkeerd gereden deelnemers"; English: "... and helped for five days repairing tubes, fixing broken chains, organising accomodation and tracking down lost participants"). From a modeling perspective, high phonetic variety may lead to good model accuracy. From a speaker perspective, shorter phrases may work to the advantage of TE speakers as the syntactic structure allows inhalation at appropriate boundaries. In other words, possible phonetic variety is easier to realize.

4. Discussion

The first objective of this study was to investigate if and how assessment models designed to predict human ratings of voice quality and speech intelligibility of tracheoesophageal (TE) speech degrade when less speech material is available for making the predictions. This question addresses the boundary conditions under which current technologies can be applied in clinical practice (cf., the studies by Clapham et al., 2014, 2015; Middag et al., 2014; Mayr et al., 2010; Stelzle et al., 2011; Windrich et al., 2008). The second objective was to check whether ignoring input features that are insufficiently supported by the speech material leads to models that are less sensitive to variations in the phonetic content of that material.

We investigated the second objective first. In fact, if we could show that ignoring input features does not degrade model performance in the case of sufficient speech material, we could restrict the study of the first objective to an analysis of the scores emerging from the best reduced-set model. Based on our earlier conjecture that the observed frequencies of infrequent phonemes of a language may differ significantly between texts of the same length, we investigated whether it is possible to reduce the sensitivity to that source of variation by prohibiting the model training to access speaker features derived from utterances of such infrequent phonemes.

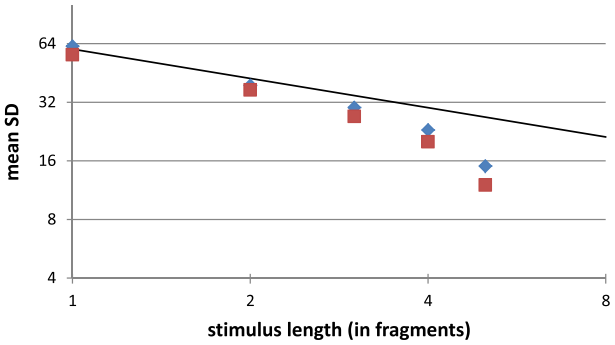

Fig. 2. Mean SD of the speaker scores as a function of the stimulus length (in text fragments). The results for voicing (diamonds) and speech intelligibility (squares) are depicted together with the trend line SD = 60/$\sqrt{\text{length}}$.

To begin with, we created five sets of speaker features that can serve as inputs to the envisioned assessment models. Using the different feature sets we then trained assessment models towards consensus ratings of two perceptual variables. Recall that these ratings can take values from 0 to 1000. We trained models that had access to one, two or three feature sets because previous studies (Clapham et al., 2014; Middag et al., 2014) had shown that combining feature sets generally results in stronger models.

We considered two conditions for the training. In the full-set condition, the models had access to all features of a feature set while in the reduced-set condition, they only had access to features referring to a sound (set) with a sufficiently high frequency of occurrence in Dutch. Note that the linear regression model training automatically determines which and how many eligible features it incorporates. Consequently, the number of model parameters is not necessarily proportional to the number of eligible features.

Comparing corresponding full-set and reduced-set models, led to the conclusion that the performance differences between both model types are not statistically significant. This means that expelling features does not hurt even when the test material (all text fragments of the speaker in this case) is long enough and matched to the length and phonetic content of the training material. In both conditions, the PMF + AMPEX and the PMF + AMPEX + ALF.PLF models attained the best voice quality models. For speech intelligibility, PMF + AMPEX + PLF model was the best (RMSE = 97.4) in the full-set condition whereas in the reduced-set condition, it was the PMF model (RMSE = 98.6). However, for both perceptual variables the best models were not statistically better than most other models. Taking all results into account, we selected the reduced-set models built on the PMF + AMPEX feature set combination as the baseline models for investigating our first objective. Note that the PMF features alone suffice to create good models, but the AMPEX features focusing on voicing and pitch stability do seem to offer a small improvement which is not so surprising given that TE speakers have difficulties in this respect (Clapham et al., 2015).

Clearly, we expect that longer test stimuli give rise to more reliable model predictions. To assess how much the phonetic **composition** of the text influences the scores we measured the SD of the scores emerging from different stimuli of a given length provided by the same speaker. To assess how stimulus **length** influences the scores, we compared the mean of these scores with the consensus score for the speaker.

The first result we can derive from Tables 2–4 is that for a stimulus length of about 50 syllables, the impact of the phonetic composition is substantial. The mean SD is equal to about 50—60% of the expected error made by the model (compare an SD of 62 to an RMSE of 122.2 and an SD of 56 to an RMSE of 98.8).

Plotting the mean SD against the stimulus length in a log-log-plot (see Fig. 2) reveals that in the beginning, the descent follows the trend that SD is inversely proportional to the square root of the stimulus length (trend line in the figure) whereas it is larger for larger stimulus lengths. As mentioned before, two stimuli composed of multiple text fragments share at least one text fragment. In fact, the longer the stimulus (in fragments) the larger the percentage of text they are sharing and the more the observed SD is an under-estimation of the SD one would have obtained with measurements on independent stimuli of the same length. The latter explains the larger descent for larger stimulus lengths. Taking everything into account, we conjecture as a second result that the impact of the phonetic composition is bound to be inversely proportional to the square root of the number of syllables in the text: it would take 200 syllables to reduce the relative impact to 25–30% of the asymptotic RMSE (obtainable with a very long text).

From Fig. 1 we conclude that the impact of the stimulus length under the assumption of equal phonetic composition is not that large: less than 15% relative for a length of 50 syllables (compare an RMSE of 154.2 to the asymptotic value of 136.8 and an RMSE of 112.9–94.7). As expected, this impact also appears to be inversely proportional to the square root of the stimulus length, as one can verify by plotting the RMSE as a function of the stimulus length in a log-log-plot.

In conclusion, voice quality and intelligibility prediction models that only have access to acoustic-based speaker features describing sufficiently frequent sounds or sound classes are robust with respect to the length of the speech material they are being tested on. When there is enough speech material available, such models are as accurate as similar models that have access to all speaker features, and, as a rule of thumb, they continue to yield accurate and stable scores for as long as the test material encompasses at least 100 syllables.

Acknowledgements

This research was supported in part by an unrestricted research grant from Atos Medical (Horby, Sweden), the Verwelius Foundation Huizen, the Netherlands) and ‘Kom op tegen Kanker’ the campaign of the Flemish League against Cancer foundation. The authors wish to thank Anna Kornman and Merel Latenstein for their assistance with the data collection.

References

Bocklet, T., Steidl, S., Nöth, E., Skodda, S., 2013. Automatic evaluation of parkinson’s speech – acoustic, prosodic and voice related cues. In: Interspeech, pp. 1149–1153.

Cicchetti, D., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess. 6 (4), 284–290.

Clapham, R., Middag, C., Hilgers, F., Martens, J.-P., van den Brekel, M., van Son, R., 2014. Developing automatic articulation, phonation and accent assessment techniques for speakers treated for advanced head and neck cancer. Speech Commun. 59, 44–54.

Clapham, R.P., van As-Brooks, C.J., van Son, R.J., Hilgers, F.J., van den Brekel, M.W., 2015. The relationship between acoustic signal typing and perceptual evaluation of tracheoesophageal voice quality for sustained vowels. J. Voice 29 (4), http://dx.doi.org/10.1016/j.jvoice.2014.10.002, 517.e23–517.e29.

Ghio, A., Revis, J., Merienne, S., Giovanni, A., 2013. Top-down mechanisms in dysphonia perception the need for blind tests. J. Voice 27 (4), 481–485.

Luyckx, K., Kloots, H., Cousse, E., Gillis, S., 2007. Klankfrequenties in het nederlands. In: Sandra, D. (Ed.), Tussen Taal, Spelling en Onderwijs. Essays bij het emeritaat van Frans Daems. Gent. Academia Press, pp. 145–154.

Mayr, S., Burkhardt, K., Schuster, M., Rogler, K., Maier, A., Iro, H., 2010. The use of automatic speech recognition showing the influence of nasality on speech intelligibility. Eur. Arch. Otorhinolaryngol. 267 (11), 1719–1725.

Middag, C., Clapham, R.P., van Son, R., Martens, J.-P., 2014. Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer. Comput Speech Lang 28 (2), 467–482.

Middag, C., Saeys, Y., Martens, J.-P., 2010. Towards an asr-free objective analysis of pathological speech. In: Proceedings of the International Conference on Spoken Language Processing, Tokyo, Japan, pp. 294–297.

Miller, N., 2013. Measuring up to speech intelligibility. Int. J. Lang. Commun. Disord. 48 (6), 601–612.

Moerman, M., Martens, J.-P., Dejonckere, P., 2015. Multidimensional assessment of strongly irregular voices such as in substitution voicing and spasmodic dysphonia: a compilation of own research. Logoped. Phoniatr. Vocol. 40 (01), 24–29.

Moerman, M., Pieters, G., Martens, J.-P., Van der Borgt, M.-J., Dejonckere, P., 2004. Objective evaluation of the quality of substitution voices. Eur. Arch. Otorhinolaryngol. Head & Neck 261 (10), 541–547.

Stelzle, F., Maier, A., Nöth, E., Bocklet, T., Knipfer, C., Schuster, M., Neukam, F., Nkenke, E., 2011. Automatic quantification of speech intelligibility in patients after treatment for oral squamous cell carcinoma. J. Oral. Maxillofac. Surg. 69 (5), 1493–1500.

Van Immerseel, L., Martens, J.-P., 1996. Pitch and voiced/unvoiced determination with an auditory model. J. Acoust. Soc. Am. 91 (6), 3511–3526.

Windrich, M., Maier, A., Kohler, R., Noth, E., Nkenke, E., Eysholdt, U., Schuster, M., 2008. Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma. Folia Phoniatr. Logop 60 (3), 151–156.