Voice Quality:
An empirical assessment & a computational model

Androutsos Dimitrios
4100522

Advanced planning and decision making
Computing Science
Department of Information and Computing Sciences
Utrecht University
Utrecht, The Netherlands

Supervisor: dr.dr. E. L. van den Broek
2nd readers: dr.ir. A.F. van der Stappen and dr. J.G. Beerends (TNO)

August 16, 2016

Acknowledgements

I would like to thank my supervisor Egon van den Broek, for the great amount of help that he provided throughout my thesis, without which it would be impossible to ever finish it. Also, I would like to thank my 2nd reader John Beerends, for providing as the database, helping as formulate the goals of this thesis and providing valuable feedback and Frank van der Stappen for his time to read and evaluate this thesis.

I would like to thank my family for supporting me both morally and financially throughout this important stage of my life.

Finally, I would like to thank all my friends who supported me in any way during my whole studies.

Abstract

Current objective assessments of speech signals show little correlation with the listener's perceived voice quality (VQ) , with their quality of experience. To remedy this omission in our knowledge on the voice, a survey was executed, including 102 listeners, who each provided their Self-Assessment Manikin(SAM) on 100 (i.e., 4× 25) speech samples of two males and two females. These samples were either high quality or degraded by pink noise, impulse noise, packet loss, or bandwidth reduction. An repeated measures analysis of variance (ANOVA) on the obtained SAM , speaker gender, and signal quality revealed that the listeners preferred one female voice and that degradations influences the SAM . The SAM was also compared with International Telecommunication Union Telecommunication standardization sector (ITU-T) 's Perceptual Objective Listening Quality Assessment (POLQA) , which showed to handle the degradations excellently; but, was unable to assess VQ adequately. To resolve POLQA 's weak spot, we developed initial computational models, founded on paralinguistic parameters solely. These models correctly predicted VQ in 87.84% (4 levels) and 70.58% (8 levels) of the cases. Unknown speaker's VQ was predicted correctly in 88.71% (4 levels) and 70.42% (8 levels) of the cases. The results of this empirical study emphasize that VQ is a complex, multidimensional construct, which is influenced by several types of common noise. Moreover, it shows that ITU-T 's POLQA can be provided with an add-on, which enables it to predict VQ as well. As such, this study provides a major step towards understanding VQ and including it in ITU-T 's standards.

Keywords: voice, speech, Quality of Experience, paralinguistics, degradations, subjective, evaluations, Perceptual Objective Listening Quality Assessment (POLQA)

I've learned that people will forget what you said, people will forget what you did, but people will never forget how you made them feel.

Maya Angelou (1928-2014)

1 Introduction

Whether or not a voice is appreciated remains hard to determine in advance, as it is still a largely unknown what factors determine voice quality. What voice features make us appreciate a voice? When these features are identified [?, ?, ?], speech recordings and real time speech transmission could be optimized and for voice-induced professions quality checks could be developed. Moreover, biometric profiles could be enriched using voice features [?, ?]. A voice-based biometric could be used to unlock a mobile phone or to access Google Home and Amazon Echo's services [?, ?].

Voice quality is considered to be part of speech quality [?], in addition to signal quality; that is, the audio quality involving degradations that may occur during audio recording or transmission, such as packet loss and pink noise. Voice quality is considered to be the paralinguistic aspect of speech. It is about "how you say something rather than what you say" [?] (p. 3). This can include many factors such as age, gender, emotion, personality and intoxication.

Over the past few decades, speech signal quality have been studied exhaustively, which resulted in many algorithms and models for a variety of purposes. For example, Beerends and Stemerdink's PSQM [?] model, its successors Perceptual Evaluation of Speech Quality (PESQ) [?, ?], and subsequently, POLQA [?, ?], predicts listeners' Mean Opinion Score (MOS) on distorted voice recordings.

A rich history exists on voice quality studies have been used to determine dysphonia [?,?]. An additional aim emerged in detecting speaker's affective state via his voice, which is considered to be part of the domain affective computing [?]: "the scientific understanding and computation of the mechanisms underlying affect and their embodiment in machines" [?]. Studies to determine how the gender and age of speakers and listeners affect the perceptual evaluations of listeners [?, ?, ?].

These studies used either sustained vowels or running speech to evaluate voice quality. Recent studies suggest that the use of running speech should be preferred, since incorporates variations in vocal characteristics, where studies using sustained vowels showed to have either a poor reliability or poor documentation [?, ?].

Voice quality has been evaluated mainly using two types of evaluators: expert or naive listeners. However, when a listener can be considered to be an expert is not well defined. For example, in [?], experts were used to predict layperson's MOS , where other studies used speech pathologists, otolaryngologists, and voice teachers as experts [?, ?]. Sofranko and Prosek [?] showed that expert's opinions among these distinct experts show a low correlation, emphasizing the absence of something like a generic expert listener.

Semantics can also significantly influence voice quality evaluations, as is shown in the field of affective computing [?, ?]. Listeners familiar with the speech sample's language, may adapt their evaluations based on semantics [?, ?]. Listeners unfamiliar with the speech sample's language, evaluate only voice quality, although a lack of understanding the speech samples might also trigger participant's agitation. Moreover, paralinguistic aspects between languages vary and can influence voice quality evaluations. For example, it has been shown that the assessment of speech accent is influenced by semantics! [?]. To mitigate this issue, the current study includes this aspect as well, including both native and non-native listeners as participants.

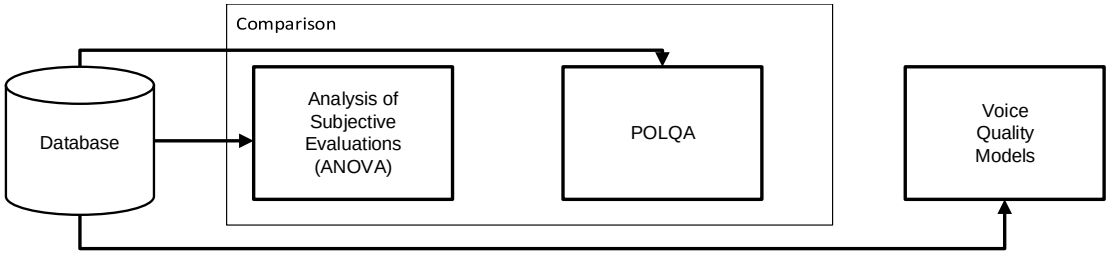


Figure 1: Schematic Overview of article's Milestones.

There are two ways to collect perceptual evaluations [**?**, **?**], including ones on speech quality [**?**]: with and without a reference sample. These two evaluations provide clearly distinct results, as has been shown over and over again (e.g., [**?**, **?**, **?**]), which can be attributed to entirely different cognitive processes: memory-based recall and perceptual discrimination. The first paradigm cannot be considered ecologically valid, as in real life practice in the big majority of cases, the situation does not concern a perceptual discrimination task. Therefore, for the current research, we adopted the former paradigm.

MOS are mainly collected via the qualppearing nterval Scale (EAIS) and Visual Analog Scale (VAS) scales, with VAS being the generally preferred scale [**?**]. For The current study, we used the SAM , which closely resembles VAS ; but, is tailored to subjective experience being used most for this purpose.

We propose to open up speech quality's decomposition and add a third dimension to it. In addition to i) signal quality, as we have defined, and ii) voice quality, iii) its affective character should be considered. We define voice quality as a complex interaction between vocal tract configuration, laryngeal anatomy, and a learned component (e.g., an accent) and its affective character as the voice, being an indirect affective signal, reflecting the speaker's neurophysiological state, which expresses emotions and moods [**?**]. Both voice quality the its affective character influence listener's quality of experience (i.e., the level of appreciation for the voice) [**?**, **?**, **?**], which makes it inherently hard to untangle. To the authors' knowledge, no studies have been reported that assess the influence of non-affective voice characteristics on the listener's quality of experience. This article aims to address this scientific lagoon by way of an large empirical study.

In the next sections, we introduce the methods of our empirical study, including the participants of the survey(Section 2.1), a description of the database used (Section 2.2), the procedure(Section 2.4), a description of the online research portal developed (Section 2.3). Next, Section 3 provides the results of the empirical study, consisting of an analysis of significant paralinguistic factors. In Section 4, the POLQA model [**?**, **?**] is evaluated in relation to the participants' subjective experienced voice quality. In Section 5 classifiers to evaluate voice quality are generated and evaluated, these classifiers can potentially be used to improve POLQA 's assessment of voice quality. Note that these research phases are also shown in Figure 1. We

end this thesis in Section 6, where we summarize the main results, discuss this research's pros and cons, and present ideas for future work.

2 Methods

2.1 Participants

102 participants answered more than 90% of the demographic questions asked when entering the online portal. Table 1 provides the descriptive statistics of these participants. Please note that 24 (23.52%) were native Dutch speakers and 61 (59.80%) were native Greek speakers, with the remaining 16 participants distributed over 9 different languages. Gender is almost perfectly balanced.

In total, the 102 participants missed 104 answers (i.e., this is 1.01% out of the total answers). To aid the forthcoming analysis, these missing answers were completed using expectation maximization, with a tolerance and convergence of 0.0001 and a maximum of 25 iterations [**?**].

2.2 Database

The database used for this research consists of 200 single channel, 16-bit, 48 kHz PCM audio files of 8 sec. length. This set of audio files consisted of speech samples of two males and two females, who all read out aloud with a neutral voice 50 sets of two subsequent Dutch sentences, separated by a brief pause. This data set was reduced to 4×25 audio files and, subsequently, these PCM files were converted to MP3 files. This limited the required network's bandwidth for the survey and ensured compatibility with all browsers.

Moreover, to assess the influence of signal quality on voice quality, the database was extended with four degradations of each undistorted audio file, namely:

- *Pink noise* is one of the widest audio degradations used. Pink noise has been used to test speech recognition algorithms [**?**], human memory [**?**] and voice quality experiments [**?**]. Moreover, it can be considered as an enhancement, since pink noise on silent parts can make the listener feel more natural [**?**]. The function addNoise of Matlab's Audio Degradation Toolbox [**?**] was used to generate pink noise, with a 48 kHz sampling frequency, zero time positions, and 64 signal-to-noise ratio.

- *Packet loss* is a common signal degradation, which mainly occurs in programs that enable voice-over-IP transmission in real-time (e.g., Skype) [**?**]. This degradation results from packet loss due to poor network conditions and methods [**?**]. Packet loss was simulated using 20 ms packets of which 10% was lost.

- *Impulse noise* contains instantaneous, impulse-like sharp sounds (e.g., clicks and pops), usually caused by electromagnetic interference, recording disks' scratches, and error prone synchronization in digital recording and communication [**?**]. A Gaussian noise mixture generator was used to generate impulsive noise [**?**] η, which was added to the speech recordings following: $x + 0.1\eta \times 10^{-1.5}$, with x being the original speech sample, using 0.05 as mixing factor per column and respectively 1.00 and 100 for the variance of the Gaussian non-impulsive and impulsive probability density functions.

Table 1: 102 participants' demographics, as included in the final database.

nationality		age		level of education		gender	
Dutch	24	18 – 25	29	elementary school (basisschool)	3	male	53
Greek	61	26 – 30	33	high school (VMBO/HAVO/VWO)	18	female	49
Indian	5	31 – 35	5	bachelor (BA/BSc)	43		
Italian	4	36 – 40	6	master (MA/MSc)	32		
Romanian	2	40 – 50	5	PhD / doctoral	6		
Japanese	1	50 – 60	13				
Chinese	1	> 60	11				
USA	1						
Tibetan	1						
Venezuelan	1						
German	1						

- *Bandwidth reduction* is a common degradation resulting from narrow bandwidth used by telecommunication companies. For this purpose a narrowband filter was implemented on the speech recordings using the Audio Degradation Toolbox [**?**]. The function applyLowpassFilter of Matlab's Audio Degradation Toolbox [**?**] was used to generate bandwidth reduction, with 8000 Hz as pass frequency, 0 Hz as stop frequency, and zero time positions.

2.3 Online portal

For the acquisition of subjective results, an online, PC and Mac compatible, portal was created. The portal was generated using PHP, HTML, JavaScript, with the subjective evaluations stored in a MySQL database.

The portal's participants were asked to listen to each of the 100 speech samples separately and evaluate it immediately after listening. The instructions provided to the participants can be found at Appendix A.
Using the SAM [**?**] (i.e., a 1–9 Likert scale used to assess a participant's emotional state) they were able to provide their evaluation (Figure 2). Participants were able to listen as often to each speech sample as they wanted. The choice to either evaluate or reproduce the speech sample appeared directly after the speech file reproduction ended. This procedure would repeat until the participant had evaluated all the 100 speech files.

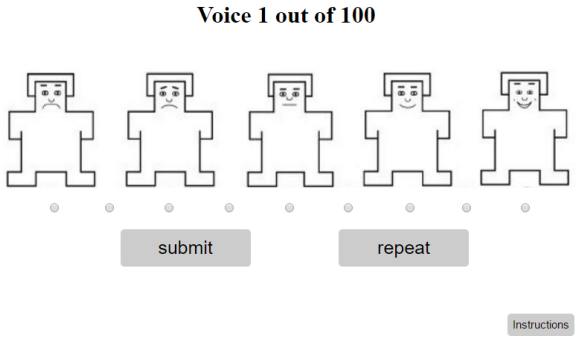


Figure 2: The survey's Self-Assessment Manikin(SAM) evaluation screen.

2.4 Procedure

As discussed in Section 2.2, the database contained degraded speech files, in addition to the original undistorted speech files. In this Section, the order in which these speech files are presented to the listeners is described.

These four degradations were applied on the complete data set of 4×24 speech samples. Together with the original data set, this resulted in $5 \times 4 \times 25$ speech samples. From these 500 speech samples, 20 ordered lists of 100 speech samples were generated using algorithm 1.

Algorithm 1 Generation of ordered lists.

1: generate *list*
2: **for** each speaker $\{f_1, f_2, m_1, m_2\}$, generate set s of his 25 shuffled speech samples **do**
3: split s into 5 subsets of 5 samples: s_0, s_1, s_2, s_3, s_4
4: apply *pink noise* to s_1
5: apply *packet loss* to s_2
6: apply *impulse noise* to s_3
7: apply *bandwidth reduction* to s_4
8: leave s_0 unchanged
9: add s at the end of *list*
10: **end for**
11: shuffle *list*
12: return *list*

3 Analysis of subjective evaluations

To determine whether or not the participants' subjective ratings unveiled effects of degradations (5), speaker gender (2), and speakers per gender (2), an ANOVA was executed. Table 2 provides the basic statistics of the participants' MOS for the independent variables just mentioned. Table 3 shows the results of the ANOVA for all factors that did have a significant influence. The results show that all factors included do have a significant impact on the subjective ratings. This includes paralinguistic effects, as can be inferred from the effects of gender and speaker.

To control for possible effects of participants' nationality, age, gender, education level (see Table 1), and audio reproduction system, complementary analysis were conducted.

In general, Dutch participants, provided higher evaluations than non-Dutch participants.

Moreover, the interaction between speaker's gender and participant's nationality showed to have a significant influence.

The effect of gender was stronger on non-Dutch than on Dutch participants $F(1, 100) = 7.717, p = 0.007, \eta^2 = 0.072$, as shown in Figure 3. Also, an interaction effect was unveiled between the participant's and the speaker's gender, $F(1, 100) = 4.302, p = 0.041, \eta^2 = 0.041$, as denoted in Figure 3. Both genders exhibit a preference for the female voice; but, females appreciate the female voice more than males do, while females appreciate the male voice less than males do.

4 POLQA's evaluation

Recently, the ITU-T has defined its third generation standard on speech quality evaluation: POLQA [**?**, **?**]. POLQA is mainly based on signal quality aspects. However, as shown in the previous section, paralinguistic features do play a significant role as well. This section assesses POLQA 's performance on both signal quality aspects and paralinguistic features. Figure 5

Table 2: Descriptive statistics of the participants' Mean Opinion Score (MOS; range 1 – 9).

	mean	SD	SE	95% CI
female 1	5.244	2.211	.093	5.060 – 5.429
female 2	4.755	2.076	.089	4.578 – 4.931
male 1	4.546	2.029	.089	4.369 – 4.722
male 2	4.587	2.035	.087	4.414 – 4.760
Dutch	5.094	2.266	.160	4.778 – 5.411
non-Dutch	4.692	2.050	.086	4.521 – 4.863
undistorted	6.080	1.946	.121	5.839 – 6.321
pink noise	4.952	1.944	.140	4.657 – 5.229
packet loss	3.519	1.905	.142	3.238 – 3.800
impulse noise	3.990	1.812	.130	3.731 – 4.248
bandwidth reduction	5.375	1.853	.121	5.135 – 5.614

Legend: SD: Standard Deviation; SE: Standard Error; and CI: Confidence Interval.

Table 3: Results of a repeated measures ANOVA on the effects of the 5 *degradations*, speaker *gender* (2), and the *speaker* per gender (2) as well as their combinations on participant's Mean Opinion Score (MOS).

degradations	gender	speaker	Specification of effect				
•			F(4,98)	=	37.354,	p < .001,	$\eta^2 = .604$
	•		F(1,101)	=	46.292,	p < .001,	$\eta^2 = .314$
		•	F(1,101)	=	19.798,	p < .001,	$\eta^2 = .164$
•	•		F(4,98)	=	9.732,	p < .001,	$\eta^2 = .284$
•		•	F(4,98)	=	10.056,	p < .001,	$\eta^2 = .291$
	•	•	F(1,101)	=	43.662,	p < .001,	$\eta^2 = .302$
•	•	•	F(4,98)	=	3.209,	p = .016,	$\eta^2 = .116$

provides a scheme of this assessment.

POLQA objective evaluations and the subjective ones have two notable differences. POLQA mainly focuses on signal distortions, so as mentioned in Section 1 it was trained by providing both a degraded and a reference sample to the listeners. In this study we focus in voice quality therefore our subjective evaluations were collected without providing a reference sample.

Additionally, POLQA 's assessment was computed for each speech file, using a scale from 1 to 5. Since, the survey's subjective ratings scale was 1 – 9, these were re-scaled to 1 – 5 (cf. [**?**]).

POLQA and participant's MOS showed high Pearson correlation for both POLQA v1.1 ($r = 0.845, p < 0.001$) and POLQA v2.4 ($r = 0.840, p < 0.001$). Subsequently, an ANOVA was executed to compare both POLQA 's results with the participant's MOS per speech sample in more detail, with speakers and degradations as between subject factors. A significant difference between the POLQA evaluations and the partic-

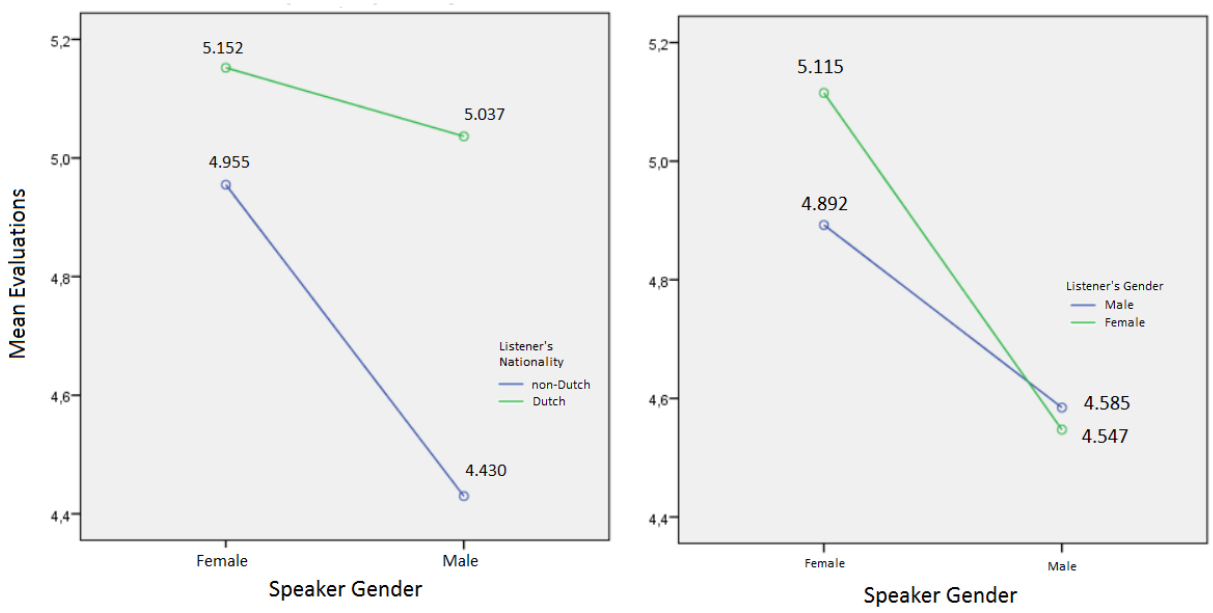


Figure 3: Left: Average valuations of Dutch and non-Dutch listeners, for speakers of different gender. Right: Average evaluations of male and female listeners, for the equivalent speakers.

ipant's MOS , also see Table 4. The difference was also significant, when the factors degradations (5), speakers (4), and their combination were included.

At first glance, the ANOVA and Pearson's correlation seem to contradict each other. However, when considering the ANOVA 's η^2 values, this can be well explained. Without including degradations, speakers and their combination as factors, the ANOVA analysis was able to explain, respectively 71.5% (v1.1) and 79.3% (v2.4) of the variance between POLQA and participant's MOS . When including degradations as factor, the explained variance rose to respectively 80.8% (v1.1) and 86.9% (v2.4) (cf. see Figures 5. Hence, POLQA is able to handle the various types of noise very well. In contrast, when including speakers as factor, the explained variance dropped sharply to respectively 29.5% (v1.1) and 27.1% (v2.4) (cf. see Figures 5. Hence, POLQA does not handle paralinguistic characteristics very well. When both degradations and speakers are both included as factors, the explained variance drops to the very low values of respectively 6.5% (v1.1) and 8.8% (v2.4). So, POLQA is a generic, robust model; but, data suggest that it can be improved or extended significantly by taking speaker's paralinguistic features into account.

5 Modeling voice quality

In this section, models specifically for voice quality are created. The process of model develop-

Table 4: The comparison of the POLQA's objective evaluations and the participant's subjective Mean Opinion Scores (MOS), taking into consideration the factors of degradations and speakers. All effect have a p<.001.

	POLQA Evaluations (v2.4)	POLQA Evaluations (v1.1)
	$F(1,480) = 1202.478$; $\eta^2 = .715$	$F(1,480) = 1843.335$; $\eta^2 = .793$
degradations	$F(4,480) = 794.363$; $\eta^2 = .869$	$F(4,480) = 503.316$; $\eta^2 = .808$
speakers	$F(3,480) = 59.394$; $\eta^2 = .271$	$F(3,480) = 67.079$; $\eta^2 = .295$
degradations & speakers	$F(12,480) = 3.863$; $\eta^2 = .088$	$F(12,480) = 2.761$; $\eta^2 = .065$

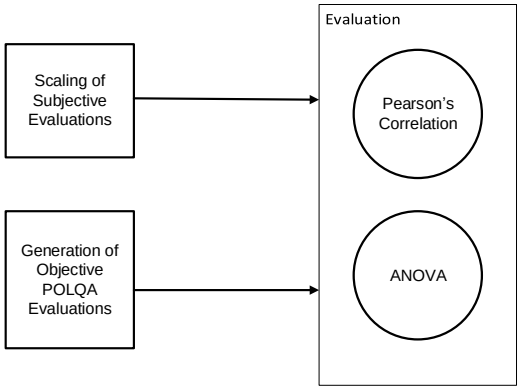


Figure 4: Schematic overview of the POLQA evaluation. Initially the POLQA objective evaluations were generated for the speech files and the subjective evaluations were scaled to 1–5. Pearson's correlation was computed for subjective - objective evaluations. Furthermore the average subjective evaluation was computed for each speech file, POLQA's objective evaluations were compared against these averages using ANOVA.

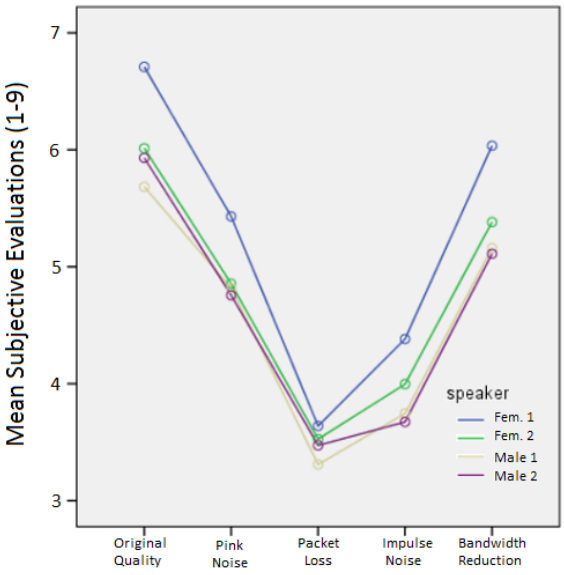
ment is shown in Figure 6. Throughout this process, only the voice's paralinguisitic features are used. Next, we will discuss each of the process building blocks.

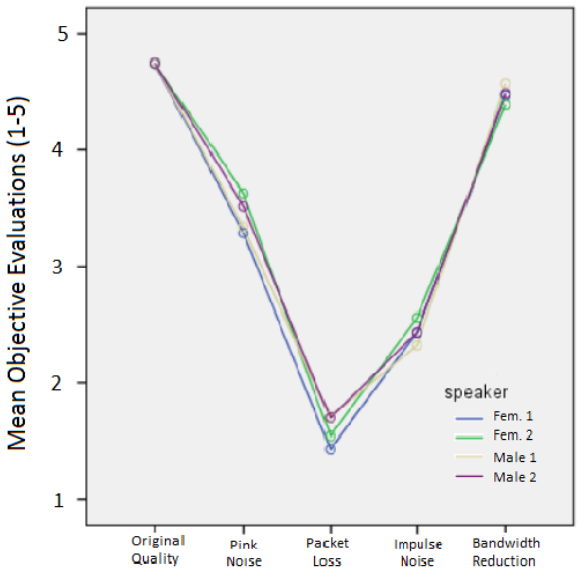
Table 5: Cross validation over random groups. The data were randomly separated in 5 groups. Training was performed excluding one group and using it for testing. Each row of data represents the R^2, R_a^2, Mean Square Error (MSE) and Root Mean Square Error (RMSE) values computed each time a group was excluded. The final line presents the mean scores achieved for each field.

Group	R^2	R_a^2	MSE	RMSE
1	.608	.574	.321	.567
2	.353	.335	.268	.518
3	.568	.543	.348	.590
4	.562	.531	.320	.565
5	.452	.429	2.752	1.659
Mean	.509	.482	.802	.780

5.1 Validation

Assume we can train a model, using a part of the available data set. This training process optimizes the classifier's parameters, such that it fits the training data. To prevent overfitting, the model's performance is validated using an independent data sample [**?**, **?**]. Here, we employ so-called cross validation, which deviates from the general validation scheme since it enables the classifier validation, without the need of an explicit validation set. As such, it optimizes the data set's use for training. Leave-One-Out Cross





	High Quality	Pink Noise	Packet Loss	Packet Loss	Bandwidth Reduction
	average(AVE) (Standard Deviation (SD))	AVE (SD)	AVE (SD)	AVE (SD)	AVE (SD)
female 1	6,74 (1,80)	5,39 (1,99)	3,68 (2,10)	4,44 (1,94)	5,98 (1,78)
female 2	5,96 (1,91)	4,85 (1,98)	3,57 (1,95)	4,07 (1,85)	5,32 (1,78)
male 1	5,74 (1,98)	4,79 (1,83)	3,31 (1,78)	3,75 (1,70)	5,15 (1,81)
male 2	5,88 (1,94)	4,77 (1,91)	3,52 (1,76)	3,71 (1,65)	5,05 (1,91)

Figure 5: Left: The mean of all subjective evaluations, for all speakers over all degradations. The table contains analytically the descriptive statistics of this figure. Right: The mean of the POLQA version 2.4 objective evaluations, for all speakers over all degradations. The subjective evaluations were collected in a 1-9 scale, while POLQA produces evaluations in a 1-5 scale. Hence, the different scales in the 2 figures.

Validation (LOOCV) is one of the most used cross validation methods to estimate estimate of the classifiers true generalization ability. As such, it provides a excellent model selection criterion.

Assume we want to assess a classifier's performance on a particular data set, containing subsets x_i with known labels c_i^l, then we can apply Algorithm 2. Such a subset can be a random sample or can consist of all data gathered of one speaker. This enables an accurate estimation of the classification error $\mathcal{E}$ on this unknown person.

All results reported in this article are determined through LOOCV . More specifically, 5-fold cross validation on random groups was performed. Additionally, 4-fold cross validation was performed to assess how the model's generalizes over the distinct speakers. For more information on cross validation, LOOCV in particular,

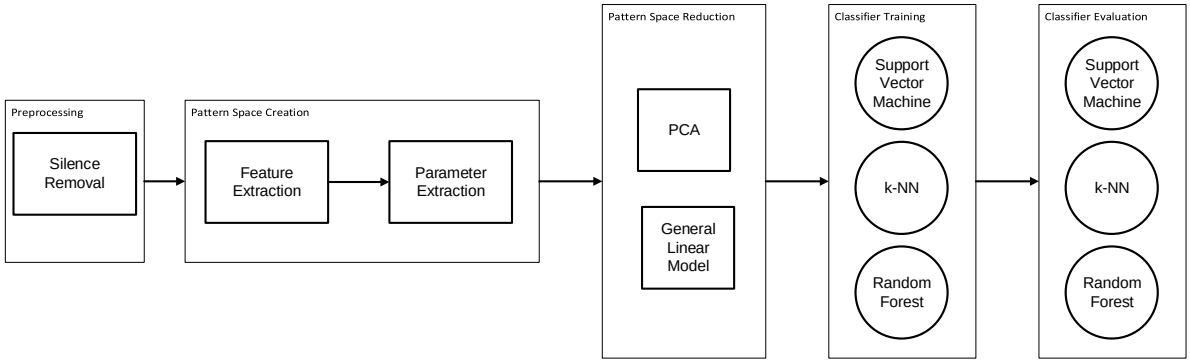


Figure 6: This is the procedure of the generation and validation of the classifiers. The pattern space was generated from the original speech files. This pattern space was reduced using two methods general linear model significancies and PCA . Using the resulted pattern spaces voice quality evaluators were trained using SVM , k-NN and RF . The performance of these models was evaluated using cross validation.

Algorithm 2 The algorithm for Leave-One-Out Cross Validation (LOOCV), adopted from [**?**].

1: **for** all i train classifier C_i with the complete data set x, except subset x_i. **do**
2: **for** all i classify subset x_i to class c_i, using classifier C_i. **do**
3: Compute C_i's average error through

$$\frac{1}{D} \operatorname*{argmax}_{c \in C} \sum_{i=0}^{D-1} \gamma(c_i, c_i^l),$$

where D is the number of subsets and $\gamma(.)$ is a Boolean function that returns 1 if $c_i = c_i^l$ and 0 otherwise.
4: **end for**
5: **end for**

we refer to [**?**].

5.2 Feature extraction

With the data gathered, we have spanned up our measurement space. Next, one type of preprocessing is needed: removal of speech's silent intervals, as these contain no signal, only noise, which could influence the feature values extracted. The following 10 features were extracted:

- Speaker's fundamental frequency of pitch (F0) [**?**, **?**];
- First formant (F1) [**?**, **?**, **?**] indicates the oral cavity's degree of closing versus opening, inversely related to vowel height;
- Second formant (F2) [**?**, **?**, **?**] indicates the tongue's back displacement versus front displacement;

Table 6: Cross validation over speakers. The data were separated in 4 groups, each one containing all but one speaker. The first column contains the speaker excluded. Each row demonstrates the R^2, R_a^2, Mean Square Error (MSE) and Root Mean Square Error (RMSE) values computed for each group. The final line presents the averages values for each field.

Speaker	R^2	R_a^2	MSE	RMSE
Female 1	.194	.170	.605	.778
Female 2	.518	.496	.382	.618
Male 1	.445	.421	.115	.339
Male 2	.572	.545	.314	.561
Mean	.432	.408	.354	.574

- Third formant (F3) [?, ?, ?], which is involved in the differentiation between between rounded and unrounded vowels (e.g., between [i] and [y]);
- Fourth formant (F4) [?, ?, ?], a higher formant, reported to significantly influence voice quality;
- Intensity [?];
- Mel Frequency Cepstral Coefficient (MFCC) [?, ?, ?]: a representation of sound's short-term power spectrum, based on a linear cosine transform of a log power spectrum on a nonlinear Mel frequency scale; and
- Harmonics-to-Noise Ratio (HNR) [?, ?, ?].
- Jitter, known to influence a voice's attractivity [?, ?, ?, ?];
- Shimmer, known to influence a voice's attractivity [?, ?, ?, ?];

Specifications on the features selection is provided in Appendix B.

From the first 8 of 10 features, the following parameters were calculated: mean, minimum (min) , maximum (max) , and SD . From jitter, the following parameters were calculated: the average absolute difference between consecutive periods, divided by the average period (jitter-local) , the average absolute difference between consecutive periods, in seconds (jitter-local-absolute) , the Relative Average Perturbation, that is the average absolute difference between a period and the average of it and its two neighbours, divided by the average period (rap) , the five-point Period Perturbation Quotient, that is the average absolute difference between a period and the average of it and its four closest neighbours, divided by the average period (ppq5) , and the average absolute difference between consecutive differences between consecutive periods, divided by the average period (jitter-ddp) . From shimmer, the following parameters were calculated: the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude (shimmer-local) , the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20 (shimmer-local-db) , the average absolute difference between consecutive differences between the amplitudes of consecutive periods (shimmer-ddp) , the 3, 5, 11 point Amplitude Perturbation Quotients, which are the average absolute difference between the amplitude of a period and the average of the amplitudes of it and 3, 5, 11 respectively closest neighbours, divided by the average (apq3, apq5, apq11). This resulted in a pattern space consisting of 42 parameters.

Next, we will discuss how the dimensionality of the pattern space was reduced.

5.3 Pattern space reduction

To reduce the dimensionality of the pattern space either feature selection or a pattern space transformation can be conducted. Multiple Linear Regression(MLR) is one of the most straight forward approaches is to conduct a feature selection. Here, we applied a MLR using a stepwise method. Two types of models were generated: an average model and a speaker model, which were cross validated as described in Section 5.1. This resulted in an adapted R^2 of .482 (Root Mean Squared Error (RMSE) = .780) and .408 (RMSE = .574) for respectively the average and the speaker model, in both cases using only 3 parameters and a constant. For the average ...

As an alternative to feature selection, a transformation of the pattern space can result in a reduction of its dimensionality. Such a transformation, defines a new set of components, using the original pattern space. For this purpose, we adopted PCA [**?**]. PCA which employs the correlations between the features of the original pattern space. Our pattern space was efficiently reduced from 42 to 9 dimensions, with 79.3% of the original pattern space variability explained. The way to select the amount of the reduced dimensions is not strictly defined [**?**](p. 10). Via an analysis of Scree plot and the eigen values, the number of components was determined. The oldest criteria is the Guttman-Kaiser Criterion [**?**, **?**], which includes components as long as their Eigen value is great than 1(1.06 for 9^{th} component). Guttman-Kaiser Criterion has received much criticism [**?**]. However, in this data seems to perform good unlinke other criteria, such as the Scree test [**?**], which chooses components based on the Scree plot (Figure 7), scree test in our pattern space would lead to fewer components being chosen, which would not explain the variability sufficiently. The implementation of PCA was performed in R using "princomp" and "factanal". Princomp scores and cor parameters set to TRUE and the rest assigned to their default values, and factanal lower was set to 0.00000005, the rotation selected being varimax, the factors selected being 9, the scores of Bartlett, and the rest of the parameters assigned their default values.

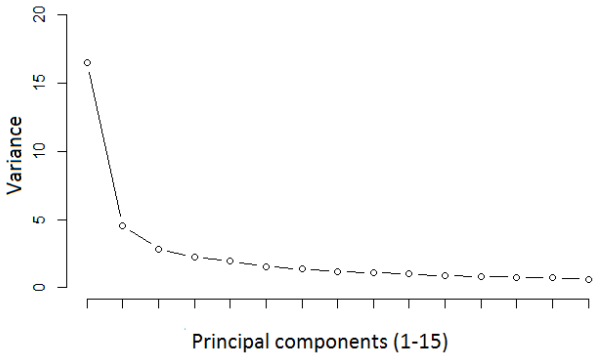


Figure 7: The Scree-plot for the 15 first components. Although the scree plot after 4^{th} component seemingly does not contribute to the variance explained, choosing 9 over 4 components, increases the cumulative proportion of variance explained from 62.2% to 79.3%.

5.4 Classifiers

To build our model, three well-known classifiers were chosen:

- The k-NN method was the 1^{st} machine learning method to be used [**?**] and is often used as baseline method, although it can also perform very well. R's library "CLASS" was used with $k = 4$ (random

groups) and $k = 7$ (speakers) and its default values.

- SVM [?] was executed using R's e1071 library, using a sigmoid kernel. The scale parameter was turned off. The kernel's need cost and gamma parameter were set to respectively 1 and 1/data dimension. For all other parameters, the library's default values were chosen.

- One of Random Forest's [?] appealing features is that it includes features selection. Hence, MLR and PCA were not used to reduce the pattern space, before executing the Random Forest. R's library randomForest was used with forest sizes 100 up to 2500, with step size 100. With each split, the number of variables used was the square root of the total number of features. For all other parameters, the library's default values were chosen.

The classifiers main results are presented in Table 7. Additionally, we note that with both k-NN and SVM , pattern space reduction using PCA resulted in better classification than when MLR was used. Moreover, for all three classifiers, results on random group cross validation were comparable with speaker cross validation.

6 Discussion

Voice quality is considered to be part of speech quality [?], in addition to signal quality; that is, the audio quality involving degradations that may occur during audio recording or transmission, such as packet loss and pink noise. Voice quality is considered to be the paralinguistic aspect of speech. It is about "how you say something rather than what you say" [?] (p. 3). This can include many factors such as age, gender, emotion, personality and intoxication, which makes it vary hard to formally specify. Consequently, it lacks a solid foundation and even its definition is debated. This is in sharp contrast with signal quality, which is exhaustively studied, including speech quality [?, ?].

Available voice quality models are founded on either single paralinguistic factors [?, ?] or are limited to features on sustained vowels, ignoring consonants, semantics, and language-related paralinguistics [?, ?]. Recent work confirmed the limitations of these studies [?] and a multidimensional perspective on voice quality is proposed. This study followed this suggestion, taking a holistic approach, exploring a plethora of paralinguistic features likely influencing voice quality, taking into account differences among speakers and gender, and assessing the influence on distortions no perceived voice quality.

The research reported here consists of three parts: i) an analysis of variance on the subjective evaluations, using listener's Mean Opinion Scores (MOS); ii) a comparison of the ITU-T 's Perceptual Objective Listening Quality Assessment (POLQA) and the MOS ; and iii) a computational model for voice quality assessment. Next, for each of these parts, the results will be summarized and discussed. Subsequently, a general discussion will be presented, including the research's pros and cons. We close with our conclusions.

First, we analyzed listeners' MOS , in relation to the speakers, gender of speakers, type of degradation. Each of these factors showed to influence listener's MOS significantly. The female voices were evaluated more positively than the male voices. However, this can be mainly attributed to one of both female voices, who's voice quality was judged significantly better than the

Table 7: The results (in %) for cross validation on both random groups and speaker (i.e., one speaker excluded when training), for Random Forest (RF), Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN), using Principal Component Analysis (PCA). to reduce the pattern space. Results for 4 and 8 levels (or classes) of voice quality are reported.

Method	RF		SVM (PCA)		k-NN (PCA)	
classes / levels	4	8	4	8	4	8
random groups	88.84	53.05	88.84	57.11	87.84	70.58
speakers	80.38	49.96	88.71	60.25	88.71	70.42

other three. This finding contradicts previous results [**?**, **?**]. Hence, a detailed analysis of the paralinguistic factors that resulted in this finding is needed. Another interesting result was that the speaker's appreciation was dependent on his/her gender and the listener's gender. Also, it should be noted that Dutch listeners' MOS was higher (5.094) than non-Dutch listeners' MOS (4.692). This can be attributed to a combination of semantics and typical Dutch paralinguistic factors, which could potentially be ignored if only vowels were examined. So far, voice quality studies focussed on sustained vowels [**?**, **?**], which is a remnant of studies related to dysphonia [**?**]. However, as this study suggest, voice quality appears to be multidimensional [**?**]; consequently, voice quality assessment founded on vowels only has a poor reliability [**?**]. Last, we note that the degradations had a major impact on the evaluations. Packet loss had the strongest impact on the evaluations, while bandwidth reduction the lightest.

Second, the subjective MOS were compared with the objective POLQA ratings (Section 4). POLQA is ITU-T 's standard for speech quality evaluation, which again showed to be very good in predicting signal quality. However, it failed in predicting voice quality and in differentiating between distinct speakers. This can be partly explained by POLQA 's timbre optimization, which eliminates voice quality differences. Although this works excellently, for signal quality assessment, it is counter productive for voice quality assessment. To remedy this conflict, the development of a complementary model is explored next.

Third, the complementary model for POLQA is developed. Paralinguistic features were selected taking the first part of this study in mind. This way, a full pattern space was generated. Subsequently, this space was reduced to aid the pattern recognition process. The resulting models were cross validated in two ways:

- Random group cross validation: Standard cross validation protocol, which resulted in a correct prediction of experienced voice quality of 87.84% and 70.58%, on respectively 4 and 8 levels of voice quality.
- Speaker cross validation: Cross validation protocol over speakers, validating on data of one unseen speaker. This resulted in a correct prediction of experienced voice quality of 88.71% and 70.42%, on respectively 4 and 8 levels of voice quality.

Using three classifiers, the models were develop: k-NN , SVM , and RF , which gave similar correct prediction rates. However, k-NN slightly

outperformed both SVM and RF . This can be explained by the relative low number of features used (i.e., < 100). Then, k-NN is capable of forming reliable neighbourhoods [**?**], where SVM and RF benefit from a higher number of features [**?**, **?**]. So, in feature research a substantially larger pattern space could be considered, to boost SVM 's and RF 's performance.

Another issue that remained not discussed so far is context. It has been shown that context influences voice's paralinguistic features. For example, in [**?**] and [**?**] it was shown that both the speech signal and biosignals are influenced by the context in which they are recorded (i.e., lab versus living room environment). Moreover, Campell and Mokhtari [**?**] showed that the same speakers voice show different paralinguistic features depending on context, more in particular depending on the social context the speaker is in. This is considered as one if not the biggest challenge in affective computing, ambient intelligence and artificial intelligence at large [**?**,**?**,**?**,**?**]. So, par excellent, this is a direction follow-up research could and should explore.

With 4 speakers included in the database, the generalizability of this study can be challenged. Complementary studies, including additional speakers or a database with speech samples of a substantially larger number of speakers would relief this issue. For example, this would enable a more detailed assessment of paralinguistic features that influence perceived voice quality. Ultimately, it could result in prototypical paralinguistic profiles of speakers and their characterization in terms of voice quality and sensitivity to noise (cf. [**?**]). Then, also speaker cross-validation would probably wield better results, since there would be more speakers with similar features to the one(s) excluded. Then, a reliable model could be constructed to evaluate a new speaker's voice quality.

Taken together, this article explored the effect of voice quality on speech quality, using an extensive listeners panel. It has been shown that voice quality significantly affects perceived speech quality. In some occasions, it affected speech quality even more than signal quality did. Moreover, many paralinguistic features showed to influence voice quality. An initial extension of ITU-T 's standard POLQA gave promising results. So, if anything, this study both illustrated the complexity of voice quality assessment as well as its feasibility in the near future.

Appendices

A Survey Instructions

Throughout this survey, you will hear four people reading sentences. Consequently, you will need to have headphones or speakers to be able to participate in this survey. **You are asked to evaluate how pleasant you feel each speech segment is. We are only interested in your honest opinion. So, there is not a right or a wrong answer in the ratings that you can give.** Each audio segment consists of two sentences.Please listen to both sentences before evaluating the audio segment. You may listen to a segment as many times as you want, using the "repeat" button. Once you have evaluated a segment, press the "submit" button to continue. The "submit" and "repeat" buttons will appear only after the audio file has been played. The time needed to complete the survey is approximately 20 minutes. You can see these instructions again at any time during the survey by pressing the "instructions" button. The survey has been tested to Google Chrome, Mozilla Firefox, Microsoft Edge, Internet Explorer, Opera and Safari. It is advised use one of these browsers. Press the "next" button to start the survey.

B Feature Extraction

In section 5.2 it was mentioned that praat [**?**] software was used to extract the voice features used in this article. In this section the praat specifications used for each voice feature extracted are described. For each audio voice recording the features were extracted by the following steps:

1. An Intensity Object (IO) was created. The minimum pitch was set 100 Hz and the time step at 0.0. The option to subtract mean was deselected, since the database voice recordings were clearly recorded, without environmental noise. Then a Textgrid (silences) Object was create. The silence threshold was set at -35 dB, the minimum allowed silent interval was set at 0.1 seconds and the minimum sound interval duration was set at 0.05 seconds. Using this object the parts that contained sound were extracted and concatenated to a file, which did not contain silent intervals greater than 0.1 second.

2. Using the IO created the mean, minimum, maximum and standard deviation of intensity were extracted.

3. A Pitch Object (PO) was created for voice recording. The timestep was set at 0.0, pitch floor 75 Hz and pitch ceiling 600 Hz. Using this object the mean, minimum, maximum and standard deviation of pitch were computed. The unit used for these features was Hertz, for the minimum and maximum the interpolation used was the parabolic one.

4. A harmonicity object was created, using the cross correlation method. The time step was set at 0.01, the minimum pitch at 75 Hz, the silence threshold at 0.1 and the periods per window

at 1.0. Using this object the minimum, maximum, mean and standard deviation HNR for each object were extracted. For minimum and maximum, the parabolic interpolation was selected.

5. A formant object was created using praat with the Burg method. The time step was set to 0.0, the window length was set at 0.025 and the max number of formants was set at 5, the maximum formant Hz was set at 5500 Hz for female speakers and 5000 for male speakers as suggested by praat and the pre-emphasis was set from 50 Hz. The mean value, the maximum value and the standard deviation for the first four formants were extracted from this object. The unit used for the measurements was Hertz and for the minimum and maximum the interpolation was set to parabolic.

6. A MFCC object was created. The number of coefficients was set at 12, the window length was set at 0.015, the time step was 0.005. The position of the first filter (mel) was 100, the distance between filters (mel) was 100 and the maximum frequency (mel) was 0.0. Using this object the MFCC matrix was generated for each audio file from which the mean, minimum, maximum and standard deviation for each object was generated.

7. A Point Process Object (PPO) was created for each audio file. To create this object the cross cross-correlation method was used and the initial audio file and its PO were combined. Then the original audio file, the PPO and the PO were combined to generate the voice-report for the audio file. The voice report was defined for the whole time of the voice recording, for pitch ranging from 75 to 600 Hz, maximum period factor 1.3, maximum amplitude factor 1.6, silence threshold 0.03 and voicing threshold 0.45. This report contains 5 jitter features (local, local-absolute, rap, ppq5, ddp), as well as, 6 shimmer features (local, local-db, apq3, apq5, apq11, dda). All these features for Jitter and Shimmer were extracted.