Modelling the influence of individual human voices on perceived quality based on ITU-T Rec. P.863

THOMAS SCHOEGJE[1], **JOHN BEERENDS**[2], *AES Fellow*, **DIMITRIOS ANDROUTSOS**[1], **EGON L. VAN DEN BROEK**[1] **AND NIELS NEUMANN**[2]

[1] *Utrecht University*
[2] *TNO P.O. Box 96800 2509JE The Hague, The Netherlands*

Speech quality is generally modelled as the audio quality of a system-enabled transfer between speaker and listener. In the absence of a reference signal however, this audio quality alone cannot explain a listener's (subjective) quality assessment. We introduce an initial speaker dependency module which can be used to extend existing audio quality models with a voice- and pronunciation agreeability model. During development, the ITU-T's Recommendation P.863 speech quality standard was used as the audio quality model. An exhaustive search through 1028 speech characteristics revealed the speaker's medium pitch during vowels as the most discriminating additional feature for P.863. The speaker dependency module was trained, tested and validated using two distinct data sets. The RSME* = 0.01 MOS of the combined model significantly outperformed P.863's RMSE* = 0.05 MOS. Moreover, we show that modelling the impact of pronunciation agreeability can further improve the speaker dependency module.

0 Introduction

Research on the quality of a system transferring speech between a speaker and a listener has traditionally focused on the *audio quality* of the system under test: is the signal before transfer (perceptually) equivalent to the signal presented to the listener (e.g. after storing and playing on an LP)? The severity of audio degradations are modelled to approximate the average human opinion when asked to evaluate the quality of the speech signal (*speech quality*). Unless asked to directly compare two signals, humans do not process speech with such a focus on the transfer medium. As a result, human evaluations made without comparison to a reference signal are biased by factors unknown to audio quality models. Gathering speech quality assessments like this is known as an Absolute Category Rating (ACR) experiment [10]. In ACR experiments subjects also take into account voice agreeability aspects and possibly pronunciation agreeability. The agreeability aspects reflect the (on average) preference for certain voices over others. This type of experiment is used because it provides quick and accurate assessment of any voice processing system, including those using signal enhancement techniques.

Speech quality can be considered as the combination of three quality aspects: audio quality, voice agreeability and pronunciation agreeability. *Audio quality* reflects distortions from the recording and reproduction system and is usually the scope of speech quality experiments. *Voice agreeability* models the characteristic and stable aspects in the voice. *Pronunciation agreeability* reflects how the speech content was expressed (due to conscious effort, the emotional state and other contextual effects). Note that all three are speech quality aspects, although the voice- and pronunciation aspects will individually be referred to as agreeability aspects.

In the development of objective measurement systems such as ITU-T's Recommendation P.863 speech quality standard [3, 4], the focus is on the audio quality. Because standard subjective quality assessments use the ACR type experiments, voice- and possibly pronunciation agreeability will also impact the perceived speech quality. These quality aspects are also unavoidable in the increasingly relevant assessment of voice-enhanced signals, where the 'degraded' output signals of a system might be preferred to the originals. Currently speech quality models (e.g. P.863 or [16]) model neither voice- nor pronunciation agreeability aspects accurately. This paper investigates the impact of all three quality aspects on the overall perceived speech quality.

Previous work on quantifying voice quality or agreeability mostly focused on specific types of outlier voice qualities, such as dysphonia [20] or creaky voices [13]. Although analysis focuses on these voice qualities, the speech characteristics used have also successfully been used to characterize typical voices (e.g. using the Harmonic-To-Noise ratio to predict speaker age[8]). An understanding of typical voice quality is also of importance in speech synthesis[9], although these works focus on emulating rather than measuring natural voice quality. Synthetic voice quality may benefit from improved voice quality models. Voice timbre is related to voice agreeability. In order to quantify the overall voice timbre speech characteristics have been introduced based on the average power density spectrum of the signal [4, 14]. Pronunciation quality or agreeability as defined here has mostly been investigated in the context of emotion recognition [7] or assigning personality traits to the speaker [17]. The term 'pronunciation quality' has also been used to score how well (non)native speakers can pronounce the speech content in a language [15].

Two subjective speech quality experiments are designed with the main goal of creating a speaker dependency module that can be used in combination with P.863. P.863 currently quantifies audio quality, and the module aims to model voice- and possibly pronunciation agreeability. In Section 1 the first speech quality experiment is introduced, focusing on quantifying a universal voice agreeability in the presence of audio quality degradations. It is followed by an ANalysis Of VAriance (ANOVA). The second experiment and its ANOVA are introduced in Section 2, and includes aspects of audio, voice- and pronunciation quality aspects. The approach to, and results of, training a speaker dependency module by extending the P.863 audio quality model is described in Section 3. This includes training a voice agreeability model on the data from the first experiment and validating it on the wider-scope data from the second experiment, as well as a further extension to an initial pronunciation agreeability model. The significance and limitations of this work are discussed in Section 4, including the importance of a more extensive pronunciation agreeability model. Finally, in the concluding Section 5 the significant findings are reiterated.

1 Human ground truth 1

The aim in the first experiment is to investigate the presence of a universal voice agreeability. In order to quantify the impact of the audio- and voice quality aspects on the overall speech quality, an experiment was set up that 1) varies the audio quality by introducing audio degradations and 2) varies the voice agreeability by having multiple speakers. The pronunciation agreeability varies minimally as speakers were instructed to always use a neutral voice. The aim is to find voice agreeability aspects which generalize. For this reason the speech content is varied as much as possible and a large number of listeners is used for the quality assessment, including many who are not native Dutch speakers.

1.1 Methodology

Speech quality is measured per individual using an ACR type scale (i.e. without a reference signal) on a 9 point MOS scale. The individual results are then averaged, resulting in Mean Opinion Scores (MOS) known as MOS-Listening Quality Subjective (MOS-LQS).These MOS-LQS results were linearly re-scaled to a 5-point scale for comparison with results from the second (5-point scale) experiment, and for comparison to P.863 predictions in terms of MOS-Listening Quality Objective (MOS-LQO 5-point scale). The methodology described in this Section largely follows ITU-T Recommendation P.800[10], as also applied in the subjective experiments used in the development of the P.863[4] audio quality model, Section 1.

First the experimental design will be introduced, followed by descriptions of the speech file preparation and the subjective quality evaluation procedure. Finally an ANOVA of the subjective results is presented.

1.1.1 Design

The audio quality was varied by using the clean reference recordings and introducing four types of audio degradations:

1. pink noise (SNR 25 dB)
2. bandwidth limitation (0-8000 Hz)
3. packet loss (losing 10% of packages at 20ms)
4. impulse noise (based on a Gaussian mixture generator available at[19] [1])

In order to vary the voice agreeability there were two male and two female speakers (aged between 30 and 63). Each of the native Dutch speakers read out 25 pairs of unique sentences, resulting in (5 audio conditions x 4 speakers x 25 sentence pairs =) 500 speech recordings. The large variation in speech content allows for a better modeling of the average voice agreeability of each of the four speakers individually.

As there are too many speech recordings to evaluate in a single session, each listener (i.e. person evaluating the speech quality) only assessed a subset of the recordings. As the experiment focuses on voice agreeability, the subset presented to each listener varies maximally in voices and speech content. As a result each listener did not hear all audio quality versions of the same file. The subsets were randomly chosen by splitting the 25 sentence pairs of each speaker into groups of 5. Each group of 5 was presented after applying one of the 5 audio quality degradation types. This results in every listener evaluating 100 speech files. In order to compensate for ordening and learning effects in the experiment, the audio files were played in one of 20 random orders to each listener.

1.1.2 Data acquisition

Four Dutch native speakers were recorded in a low-noise anechoic room with high-quality equipment. Speech was pre-processed to be super wideband (40 Hz - 14 kHz), and saved as 48 kHz PCM files. The playback levels of

[1] The impulse noise was generated using default parameter values, adding the noise mixture to each signal as $+10^{-2.5}$*noise*

the recordings were all calibrated following ITU-T Recommendation P.56[11]. The audio quality in terms of predicted MOS of the clean recordings, as estimated by P.863, lies between 4.70 and 4.75, the theoretical maximum.

1.1.3 Participants & Procedure

102 listeners gave their evaluations in an online audio experiment, including 53 males and 49 female between the ages 18 to 68. More than two thirds of the listeners (70) were not native Dutch speakers, and a majority of those did not speak the language of the speech content.

As there is less control over the playback situation in an online experiment, the participants were given the option to replay audio files and choose their own pace (with an expected session of 15 minutes). The playback equipment is unknown; but, consistent within listeners, and it resembles performance in practical applications.

1.2 Results

A repeated measures ANOVA is used on this experiment data to quantify the impact of audio quality and voice agreeability on speech quality. The experiment is designed with 4 speakers, 5 audio quality degradation types: clean, pink noise, narrowband, packet loss, impulse noise. As the gender of the speaker may also be an important aspect of voice agreeability, this 5x4 design is modeled as 5x(2x2) by introducing the variable *speaker per gender*. Because the same listeners evaluated a subset of the speech files which are tested under multiple conditions, the repeated ANOVA variant is performed. This compensates for the variability of the individual differences within subjects.

The results of the ANOVA are are provided in Table 1. All audio- and voice quality aspects investigated were significant. Audio quality proves to have the strongest effect, followed by gender and speaker per gender, both representing voice agreeability aspects. Females voices scored higher than male voices. The largest difference between listeners who were native Dutch speakers and other listeners was that native speakers showed a bias of a quarter of a point in MOS values on the 5-point scale. This could indicate a language familiarity effect.

Audio- and voice quality aspects both proved to have a significant impact on the overall speech quality for both those listeners who speak the Dutch language and those who do not. A second experiment is required to quantify the impact of pronunciation agreeability on speech quality.

2 Human ground truth 2

In order to quantify the impact of audio-, voice- *and* pronunciation quality aspects on the overall speech quality an experiment was set up that 1) varies the audio quality by introducing audio degradations, 2) varies the voice agreeability by using multiple speakers and 3) varies the pronunciation agreeability by having speakers pronounce the speech content in two different ways. These pronunciations will be indicated by the labels non-aroused and aroused.

Table 1. Results of a repeated measures ANOVA on the effects of the audio quality conditions (Q, 5) x speaker gender (G, 2) x speaker per gender (S, 2) on the MOS values. Indicated are the variables tested and their effects, with $p < .001$ in all cases.

Q	G	S	specification of effect			
•			$F(4,98)$	=	37.35	$\eta_p^2 = .60$
	•		$F(1,101)$	=	46.29	$\eta_p^2 = .31$
		•	$F(1,101)$	=	19.80	$\eta_p^2 = .16$
•	•		$F(4,98)$	=	9.73	$\eta_p^2 = .28$
•		•	$F(4,98)$	=	10.06	$\eta_p^2 = .29$
	•	•	$F(1,101)$	=	43.66	$\eta_p^2 = .30$

2.1 Methodology

This experiment is again of the ACR-type and results are measured on a 5-point MOS scale. This subjective test follows ITU-T Recommendation P.800 [10] as also applied in the subjective experiments used in the development of the P.863 [4] audio quality model, Section 1.

First the experimental design will be introduced, followed by descriptions of the speech file preparation and the subjective quality evaluation procedure. Finally, an ANOVA of the subjective results is presented.

2.1.1 Design

The audio quality was varied by using the clean reference recordings and introducing two types of audio degradations:

1. pink noise (SNR 25 dB, the same as in the voice agreeability experiment),
2. bandwidth limitation (40-3700 Hz, as opposed to 0-8000 Hz in the voice agreeability experiment)

In order to vary the voice agreeability there were 10 different native Dutch speakers, each reading out the same two sentences. There were 5 male and 5 female speakers aged from 20 to 63.

The pronunciation agreeability was varied by instructing speakers to read out the sentences in two different manners. First at a calm pace, with slightly reduced intonation and at a controlled volume (non-aroused) and then at a natural pace, with more intonation and at slightly more volume changes (aroused).

The speech content consists of four sentence pairs, resulting in (3 audio quality levels x 10 speakers x 2 pronunciations x 4 sentence pairs =) 240 speech files. These speech files were played in a unique, different, random order to each of the (native Dutch speaking) listeners. This compensates for ordening and learning effects. All listeners judged the speech quality of all speech files.

2.1.2 Data acquisition

The files resulting from 10 native speakers were first recorded and prepared as specified in the description of the first experiment. The quality of the clean reference record-

Table 2. Results of an ANOVA on the effects of the audio quality conditions (Q, 5) x speaker gender (G, 2) x speaker per gender (S, 2) x pronunciation agreeability (P, 2) on the MOS values. Indicated are the variables tested and their effects, with $p < .001$ in all cases except where $p = .006$.

Q	G	S	P	specification of effect			
•				$F(2,180)$	=	1191.72	$\eta_p^2 = .93$
	•			$F(1,180)$	=	53.18	$\eta_p^2 = .23$
		•		$F(4,180)$	=	12.53	$\eta_p^2 = .22$
			•	$F(1,180)$	=	218.64	$\eta_p^2 = .55$
•	•			$F(4,180)$	=	11.20	$\eta_p^2 = .20$
•		•		$F(8,180)$	=	2.77	$\eta_p^2 = .11$
•	•			$F(4,180)$	=	11.42	$\eta_p^2 = .20$
	•	•	•	$F(4,180)$	=	24.86	$\eta_p^2 = .36$

ings of this second experiment was then optimized by suppressing minor residues of background noise (e.g. breath of the speaker and system noise). This further increases the separation of the audio versus voice- and pronunciation agreeability aspects. Noise suppression was performed following [2], where manual noise suppression in silent intervals slightly improved quality scores over any available software. The audio quality of the clean reference files in terms of predicted MOS, as estimated by P.863, again lies between 4.70 and 4.75, the theoretical maximum.

2.1.3 Participants & Procedure

The quality assessment procedure follows the one applied in the subjective experiments used in the development of P.863. The experiment was performed at various low-noise low-reverb locations with high-quality diffuse-field equalized headphones. The speech was played back at a nominal sound pressure level in the acoustical domain of 73 dB at the Ear Reference Point [4], Section 1. The speech content was played in four sessions of 60 sentence pairs with three seconds of decision time between each pair. Between each session there was a small break (30 seconds) with a longer break halfway through the experiment. This takes around 40 minutes total.

25 listeners listeners assessed the speech quality, including 16 males and 9 females between the ages 18 to 82.

2.2 Results

The ANOVA on the second experiment indicates the impact of the various speech quality aspects on the overall speech quality. The experiment was designed to vary audio quality (clean reference, narrowband, pink noise), voice agreeability (10 speakers) and pronunciation agreeability (aroused, non-aroused). As the gender of the speaker may also be an important aspect of voice agreeability, the 3x10x2 design is modeled as 3x(2x5)x2 by introducing the variable *speaker per gender*. As each listener evaluated each file a regular type of ANOVA was performed.

All significant speech quality factors with an estimated variance (while controlling for other predictors) $\eta_p^2 \geq 0.1$ on the MOS are summarized in Table 2. All audio-, voice- and pronunciation quality aspects investigated were significant. A detailed analysis even showed that the worst clean reference files (i.e. MOS 2.35, 2.65) were judged to be of lower quality than the best degraded files (i.e. MOS 2.96, 2.92 for narrowband and MOS 3.27, 3.23 for pink noise). Audio quality proved again to have the strongest effect, and pronunciation agreeability proved to have a larger estimated effect size than either variable related to voice agreeability (gender and speaker per gender). There was a strong preference for the arousal pronunciation. speaker gender was again more important than speaker per gender, with females voices scoring higher than male voices.

A post-hoc Tukey test on audio quality shows its effect can be explained by the large difference between the scores of the clean reference, pink noise, and narrow band degraded files, with the difference between the latter two barely significant at $p = 0.048$. A post-hoc Tukey test on each of the 10 voices confirms that almost every voice significantly differs from at least two other voices, confirming the importance of individual voice agreeability.

The audio quality had a different impact on the different voices, and especially the difference in impact on male and female voices was significant. Whereas both degradations were similar for male speakers, the narrow band degradation effected the female speakers more than the pink noise degradation. This effect is related to the power spectral density differences between male and female speakers. Finally the pronunciation agreeability affected the voices differently, as the speaker's level of arousal differed significantly. This can also be seen as the duration ratio between the non-aroused and aroused conditions, which varied between 1.00 and 1.10.

The ANOVA shows that voice- and pronunciation agreeability aspects have a significant impact on the overall speech quality that is so large that the clean reference files with the lowest speech quality scored lower than the best rated files containing either audio degradation. The impact of speaker dependency can thus be of the same order as a 40-3700 Hz bandwidth limitation or or 25 dB SNR pink noise. It is expected that P.863 can be improved by taking these aspects into account.

3 Modeling based on ITU-T Rec. P.863

In contrast to audio quality, voice- and pronunciation agreeability aspects are not well understood. P.863 is focused on the audio quality. Many speech characteristics are computed based on the recorded speech files, which may quantify voice- and pronunciation agreeability. Based on these, a speaker dependency module is trained which can be used to extend an audio quality model such as P.863.

In modeling speech quality it is a priority to avoid overfitting on the training dataset. Every experiment should be considered to have a very limited context compared to all possible speech that should be modelled. In order to get a stable voice agreeability model, the training and wider scope validation are each carried out on one of the ground truth datasets from two distinct experiments (see Sections 2 and 3). The data from the first experiment has more variation in audio quality, less variation in voice agreeability and no variation in the pronunciation agreeability. Because of this a general, stable model extension can be developed by using only the first experiment in the model training and validated in a wider scope, using the data from the second experiment.

3.1 Feature extraction

1028 speech descriptors were derived from the speech signals, using the speech analysis software packages VoiceSauce 1.31 [18] and Praat 6.0.35 [6], using the packages default parameter settings. The only exception is the pitch estimation in VoiceSauce, which was done using Praat's algorithm (which is based on auto-correlation[5]). This was done in order to use the same pitch algorithm for all pitch-based characteristics. Although pitch extraction algorithms were not compared here in terms of accuracy and robustness, the Praat algorithm attained competitive accuracy in previous works [1].

The VoiceSauce parameters were computed only for voiced intervals, as they are invalid in unvoiced intervals. The Praat scripts computed four versions of each descriptor, which were using the non-silent frames from one of the following pre-processed speech recordings:

1. unaltered
2. with unvoiced parts silenced
3. with voiced parts silenced
4. with everything except the vowels removed

A complete overview of the speech descriptors investigated can be found in Table 3.

3.2 Model selection

70% of the ground truth 1 data set is used for training the combined audio- and voice quality model. The remaining 30% is used for testing.

Because any speech experiment represents a limited speaking context, the degrees of freedom of the module should be limited to prevent overfitting. Especially given the number of possible speech characteristics and dataset sizes the curse of dimensionality is significant. Discovering the best candidate models is approached by finding one or two speech characteristics that most accurately predict the subjective voice agreeability aspects while the P.863 predicted MOS is used for quantifying all audio quality aspects. This results in combined models with two or three speech characteristics of which the P.863 predicted MOS is the main indicator. Using four speech descriptors does not lead to a significantly more accurate model. For each candidate model, a multiple linear regression model is trained and tested on subsets of the first voice agreeability dataset. Each candidate model trained on the 70% training data by minimizing the Root Mean Squared Error* (RMSE*). This extension on the RMSE error uses the 95% confidence interval of the MOS values to weight the reliability of each MOS value. This is the standard error measure used in the P.863 development, and is fully described and motivated in Appendix I.4 of ITU-T Recommendation P.863 [12]. Next, the candidate models are ranked by their RMSE* scores on the 30% testing data.

After developing a combined audio- and voice quality model, a further extension is developed. This extension focuses on quantifying the impact of pronunciation agreeability. The best combined audio- and voice quality model is further extended using only the second ground truth database. Modeling this second extension follows the same methodology as of the voice agreeability model. However, as only one data set is available on pronunciation agreeability, the results cannot be validated in a wider context. For this further extension the ground truth data set 2 is also split into 70% for training, and 30% is used for testing.

3.3 Validation

3.3.1 Human ground truth 1: Voice Agreeability

For testing on the training subset of the ground truth 1 dataset, results are shown on the 30% testing data. Figure 1 shows the P.863 predicted MOS, without a voice agreeability model, compared to the subjective MOS. The results correlate fairly well due to the variety of audio quality degradations in the voice agreeability experiment, although worse than usually due to the voice agreeability aspects. Each audio quality degradation is visible as a cluster and the cluster with the clear reference files all scoring near the maximum theoretical audio quality score in P.863.

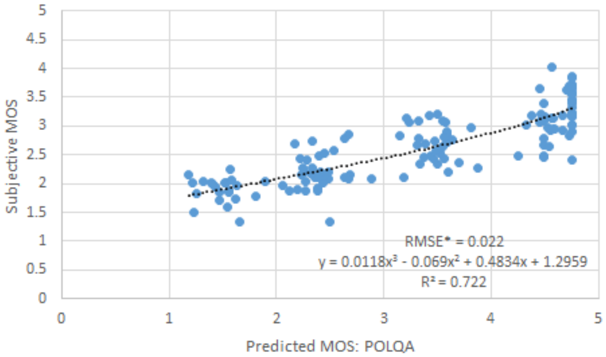

Fig. 1. Predicting the MOS on the first (training) dataset with one variable: P.863-MOS.

The best combined audio- and voice quality model using only one additional speech descriptor uses the P.863 predicted MOS and the mean location of formant 4 in Hz (vowel pre-processed). Compared to other formants the 4th formant shows less variance, and it appears to quantify the overall voice agreeability aspect. During the vowels the power density of the signal most strongly exhibits certain characteristics of the voice, such as the formants. As shown in Figure 2, this combined model no longer exhibits the strong clustering tendency that was seen with P.863. It explains the differences within the clusters, reducing the RMSE* from 0.02 to 0. These error values are low due to the large confidence intervals in the subjective results of the experiment.

Perceptual dimension	Features	Description	Parameters
Audio quality	P.863 prediction	Perceptual audio quality model	
Pitch	Fundamental frequency f0	f0 statistics	quantiles, std dev, min, max; in mel, semitones, ERB
	Mean absolute pitch slope	Slope of pitch over audio file	per sentence, average sentence, full audio file; Hz, semitones
Loudness	Intensity	Time-domain amplitude	quantiles, avg, std dev, min, max, sum, avg decrease when falling, avg increase when rising, fall/rise quotient
	Vocal effort	Strength of Excitation	
Timbre	Formants	Formant F1-F5, F1 x F2, F2 x F3 locations and bandwidths	quantiles, avg, std dev; bark, Hz
	MFCC coefficients	Representation of audio power spectrum	avg, std dev
	Harmonics	Harmonics H1-5 amplitudes	quantiles, avg, std dev, max and long-term stability[a] ..
		Harmonics A1-A3 nearest formants F1-F3	.. (un)compensated for formants, normalized, absolute
		Harmonics nearest 2000Hz and 5000Hz, H2K, H5K	see above
	Bark band values	Pitch scaled to perception	quantiles, avg, std dev; absolute, normalized
	Spectral centroids	Relative spectral energy distribution	P.863's bark timbre, Soft Phonation Index, Voice Turbulence Index, Spectral Emphasis [b]
	Spectral stability	Variation in spectral in overlapping 1 second time-windows	quantiles, avg, std dev
Vocal stability	Jitter	Cycle to cycle pitch perturbation	avg, sum, rap, ppq5, dda
	Shimmer	Cycle to cycle amplitude perturbation	avg, avg in dB, apq3, apq5, apq11, ddp
	Mean autocorrelation	Correlation of sequential pitch periods	
	Harmonics-To-Noise ratio	Ratio of signal periodicity to signal noise	avg, local std dev, max local, min local; on full spectrum and bandwidth limited to 500, 1500 and 2500 Hz
	Voice breaks	Breaks in sustained voiced speech	number of breaks, ratio unsustained per voiced pitch frames
	Cepstral Peak Prominence	Unexpected peak prominence in cepstral domain	
	Subharmonic To Harmonic ratio	Amplitude ratio between subharmonics and harmonics	
Speech rate	Absolute duration	File length	
	Duration ratios	Relative total length of intervals	voiced/unvoiced frames, voiced/total frames
Combination	Glottal pulses	Glottal closures during speech production	pulse count
	Glottal pulse periods	Periods between glottal pulses	period count, mean duration, duration std dev
	Conditional statistics	Averages per sentence, per speaker, overall	all descriptors where applicable (e.g., F1-F5 locations)
	Arousal quotient	Relative length of a file compared to its (non)arousal counterpart	
	Gender	Binary value	

Abbreviations: Equivalent Rectangular Bandwidth (ERB), Hertz (Hz), Mean Frequency Cepstral Coefficients (MFCC), deciBel (dB), x-point Period Perturbation Quotient (ppqx), x-point Amplitude Perturbation Quotient (apqx), Difference of Differences of Periods (ddp), Relative Average Perturbation (rap)

[a] Average difference of overlapping 1 second time-windows.

[b] These weight the relative intensity in the following frequency ranges: Bark bands 2-7 / Bark bands 5-12; 70-1600Hz / 1600-4500Hz; 40-4000 Hz / 50-7000 Hz; (0 through 1.5 * f0 Hz) / full signal

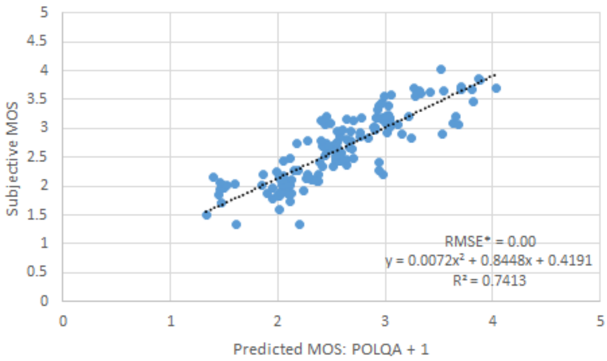

Fig. 2. Predicting the MOS on the first (training) dataset with two variables: P.863-MOS + 0.002 mean of formant 4 in Hz (vowel pre-processed.

The best combined model candidates with two additional descriptors usually combine descriptors that did well in the simpler models. The simpler model is stronger however, as the RMSE* score is only marginally improved from 0 and the top model includes a speech characteristic which overfits the audio degradations.

3.3.2 Human ground truth 2: Voice Agreeability

The resulting combined model of training on only the ground truth 1 data is now presented on the (complete) ground truth 2 dataset. Figure 3 shows the P.863 predicted MOS compared to the subjective MOS. As this experiment included more voice- and pronunciation agreeability effects, the P.863 predictions are not as accurate as normally observed in speech quality experiments. Both audio degradations affected the MOS values with similar strength and, thus, there are only two clusters. There is one degraded audio quality cluster (bandwidth 40-3700 Hz and SNR 25 dB) and a clean reference cluster near the optimal audio quality.

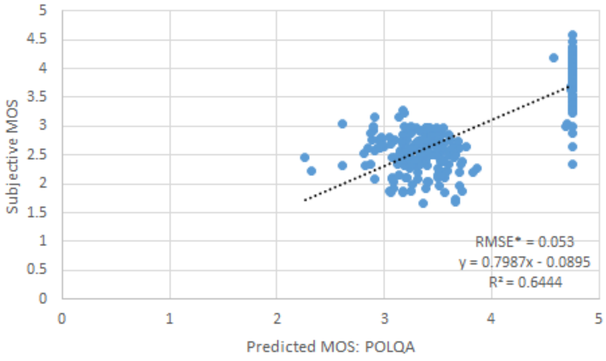

Fig. 3. Predicting the MOS on the second (testing) dataset with one variable: P.863-MOS.

When the combined audio- and voice quality models are validated on the ground truth data 2, it becomes evident that the formant 4 characteristic, which previously performed best, does not generalize. Other candidate models outperform it because the bandwidth limitation (40Hz-3700Hz) significantly affects the formant 4 values which are otherwise located near the 4000Hz. Instead of the formant 4 characteristic, the median pitch (vowel pre-processed) characteristic is chosen. This is the best candidate model that does generalize across both experiments, as the speaker-dependency module should also be accurate for medium-quality systems. Using the combined model of P.863 with this pitch characteristic the RMSE* drops significantly on the second dataset, from 0.05 to 0.01 (see Figure 4).

Pitch is related to both voice agreeability and pronunciation agreeability. Analysis shows that higher pitched voices are preferred, which is consistent with the result that female voices are preferred over male voices. Pitch is also useful in quantifying pronunciation agreeability. Analysis showed that an increase in arousal is usually accompanied by an increase in both pitch and the speech quality.

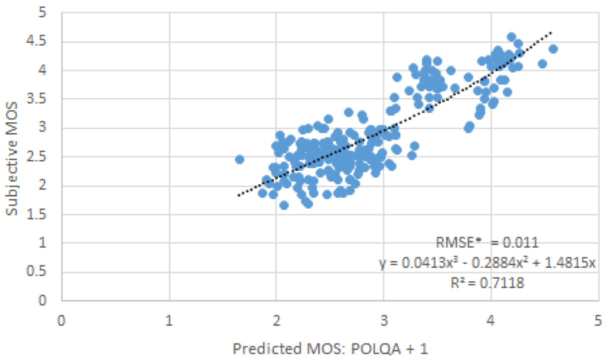

Fig. 4. Predicting the MOS on the second (testing) dataset with two variables: P.863-MOS + 0.011 median pitch in Hz (vowel pre-processed).

The large increase in model accuracy is obtained without any training on the data of the second experiment. As these experiments differ significantly in experimental scope, it shows that the voice agreeability model can be expected to have a stable behavior. Even more so as the first experiment contained many ratings from listeners that did not speak the language.A further extension of speaker dependency may be possible by training on data of the second experiment, which also varies the pronunciation agreeability.

3.3.3 Human ground truth 2: Pronunciation Agreeability

The best combined audio- and voice quality model is further extended with one characteristic in an initial attempt to quantify pronunciation agreeability. After training on 70% of the ground truth 2 dataset the best candidate already overfits the audio degradations present. The first model that does not overfit it uses a speech characteristic which roughly estimates the variation in the first two formants. It is obtained by multiplying the locations of the first two formants in Hz and taking the standard deviation (without pre-processing). The average locations of the these formants rarely exceed 2000 Hz and hence only very poor bandwidth limitations will affect this. This improves the RMSE* from 0.02 to 0, but is not expected to generalize. More experimental data is required for extending the audio- and voice quality model towards an audio-, voice- and pronunciation quality model.

4 Discussion

When assessing low-quality recording/reproduction systems, audio quality dominates the perceived subjective speech quality. When assessing high-quality recording/reproduction systems, the voice agreeability itself becomes an important contributing factor to the perceived overall speech quality when no ideal reference signal is provided.

In this paper, we investigated the effect of voice- and pronunciation agreeability in medium to high-quality voice systems (pink noise with SNR of 25 dB or better, bandwidth 40-3700 Hz or better). In these conditions, the combined effect of voice- and pronunciation agreeability is of the same order as of audio quality degradations. The best bandwidth degraded and noise degraded speech files obtained MOS values over 3.0, while the worst clean reference speech files obtained a MOS under 3.0. This shows that even in medium-quality P.863 assessments one should take into account voice- and pronunciation agreeability effects.

It has been clearly shown that both voice- and pronunciation agreeability effects contribute significantly to the overall speech quality. Two distinct experiments have been executed. These provided two data sets, used to develop a first stable speaker-dependency module to function in conjunction to the ITU-T Recommendation P.863 audio quality model. This module includes voice agreeability aspects. The combined model uses a linear combination of the P.863 predicted MOS and median pitch frequency (vowel pre-processed). The combined model significantly outperformed the P.863 audio-quality model. Subsequently, an initial pronunciation agreeability extension for the module was introduced by adding a rough measure of the degree of variation in the vowel space, although this model could not be properly validated yet.

New sets of validation data are necessary to validate the current speaker dependency model's robustness, in particular with respect to pronunciation agreeability. Preferably, these should have relative small confidence intervals. The two data sets for the current research did have confidence intervals larger than normally obtained for similar research. In particular, the second ground truth data set showed the largest confidence intervals. Listeners were uncertain about their judgments due to the variation in pronunciation agreeability. An increased number of subjects would help decrease the confidence interval.

As the speech characteristics should be valid for a wide range of conditions, other methods of varying the voice agreeability may be considered. Voice agreeability was only varied within both experiments by varying the speakers. An alternative is to quantify and modify the voice timbre and to then compare multiple timbres for the same voice in a subjective experiment. Quantification and manipulation of voice timbre is possible using the power density spectrum [14].

The speaker dependency's voice agreeability aspect proved stable across experiments, even though an ANOVA determined that listeners assess male and female voices differently. Further investigation is required to confirm whether this is because male and female voices are evaluated differently or whether this follows from the pure signal characteristics in the recorded speech. Similarly the interaction effects between voice agreeability and audio quality could be explored further. However, such extensions increase the degrees of freedom and, consequently, the likelihood of over-fitting the data.

Finally, it should be noted that all results were based on Dutch speech content with non-deviant voice- and pronunciation types. Although the voice agreeability appears to be stable whether listeners speak the language of the content or not, it is unclear whether or not voice agreeability generalizes over languages.

The speaker dependency module offers a modest addition to the P.863 audio quality model; but, with excellent results on two distinct, challenging, ground truth data sets. Voice- and pronunciation agreeability aspects are unavoidable biases during Absolute Category Rating testing and when testing an ever-increasing class of systems which apply voice enhancements. As such the speaker dependency model introduced is the first step in a further improvement of objective speech quality models, such as P.863.

5 REFERENCES

[1] O. Babacan, T. Drugman, N. d'Alessandro, N. Henrich, and T. Dutoit. A comparative study of pitch extraction algorithms on a large variety of singing sounds. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7815–7819. IEEE, 2013.

[2] J. Beerends and I. Beerends. On the assessment of high-quality voice recordings including voice postprocessing. *Journal of the Audio Engineering Society*, 63(3):174–183, 2015.

[3] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl. Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part I - temporal alignment. *Journal of the Audio Engineering Society*, 61(6):366–384, 2013.

[4] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl. Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part II-perceptual model. *Journal of the Audio Engineering Society*, 61(6):385–402, 2013.

[5] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the institute of phonetic sciences*, volume 17, pages 97–110. Amsterdam, 1993.

[6] P. P. G. Boersma et al. Praat, a system for doing phonetics by computer. *Glot international*, 5, 2002.

[7] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[8] M. M. Gorham-Rowan and J. Laures-Gore. Acoustic-perceptual correlates of voice quality in elderly men and women. *Journal of communication disorders*, 39(3):171–184, 2006.

[9] A. J. Hunt and A. W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pages 373–376. IEEE, 1996.

[10] ITU-T Recommendation. P.800: Methods for subjective determination of transmission quality. *International Telecommunication Union*, 1996.

[11] ITU-T Recommendation. P.56: Objective measurement of active speech level. *International Telecommunication Union*, 2011.

[12] ITU-T Recommendation. P.863: Perceptual objective listening quality assessment. *International Telecommunication Union*, 2011.

[13] D. H. Klatt and L. C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *the Journal of the Acoustical Society of America*, 87(2):820–857, 1990.

[14] S. Möller, A. Raake, and M. Wältermann. The sound character space of spectrally distorted telephone speech and its impact on quality. In *Audio Engineering Society Convention 124*. Audio Engineering Society, 2008.

[15] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub. Automatic scoring of pronunciation quality. *Speech communication*, 30(2):83–93, 2000.

[16] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 2, pages 749–752. IEEE, 2001.

[17] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, et al. A survey on perceived speaker traits: personality, likability, pathology, and the first challenge. *Computer Speech & Language*, 29(1):100–131, 2015.

[18] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu. Voicesauce: A program for voice analysis. *Energy*, 1(H2):H1–A1, 2010.

[19] C. Tsimenidis. Gaussian noise mixture generator. [Online; accessed January 27, 2016].

[20] F. L. Wuyts, M. S. De Bodt, G. Molenberghs, M. Remacle, L. Heylen, B. Millet, K. Van Lierde, J. Raes, and P. H. Van de Heyning. The dysphonia severity indexan objective measure of vocal quality based on a multiparameter approach. *Journal of Speech, Language, and Hearing Research*, 43(3):796–809, 2000.

THE AUTHORS