Evaluation of the ITU-T P.563 standard as an objective enhanced speech quality metric for hearing aid users

João F. Santos[1], Vijay Parsa[2], Susan Scollie[2], Tiago H. Falk[1]

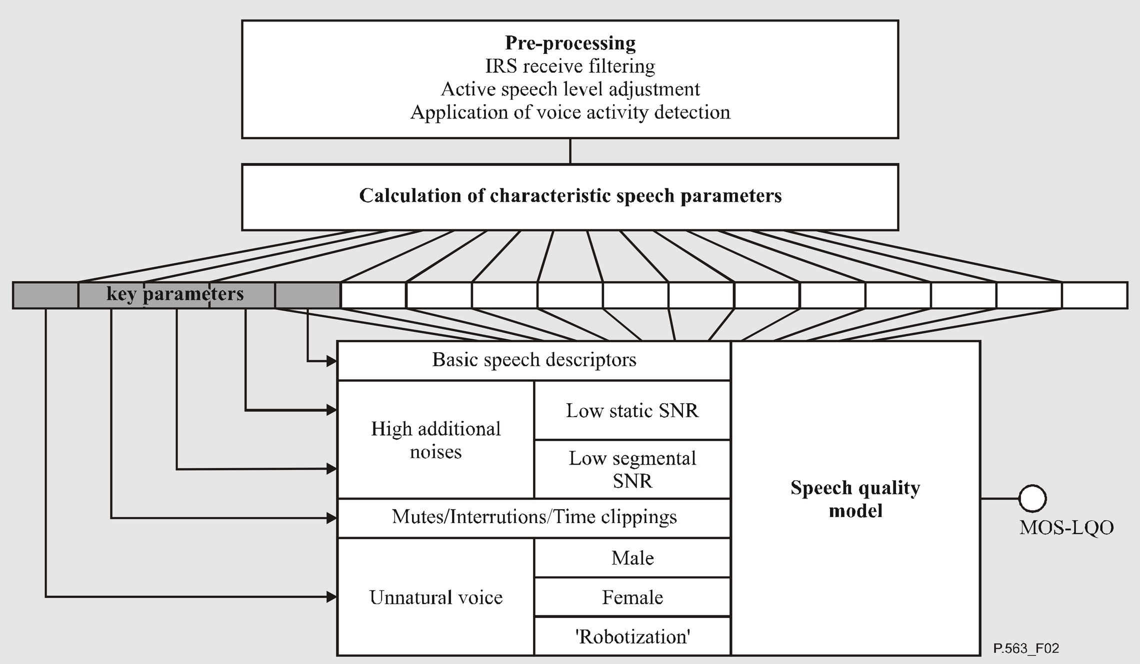
[1] INRS-EMT, University of Quebec, Montreal, QC, Canada

[2] University of Western Ontario, National Centre for Audiology, London, ON, Canada

Motivation and Goals

- P.563 is an ITU-T standardized single-sided speech quality metric for narrow-band telephony applications [1].
- No similar standards exist for hearing aid (HA) users.
- Objective speech quality assessment is less laborious and time-consuming than subjective testing.
- We investigated the performance of P.563 and its internal parameters as a tool for assessing speech quality for HA users.

P.563 internal features

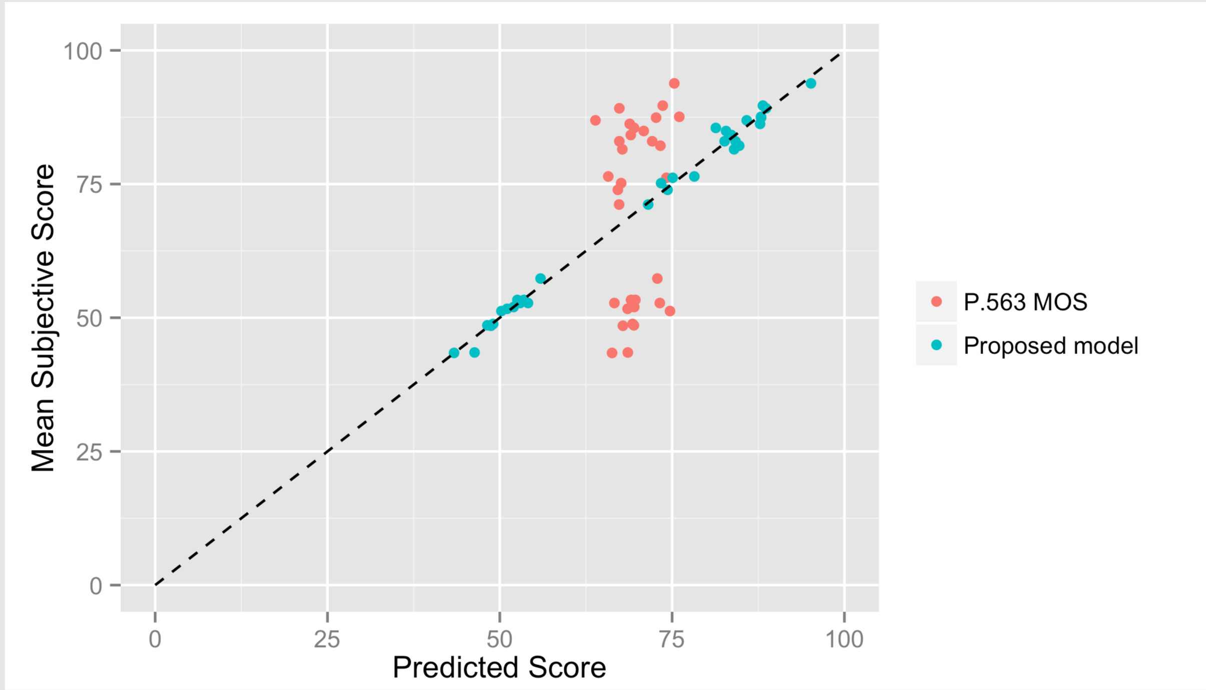


- P.563 is based on 43 internal features, related to six dominantly distortion classes (background noise, signal interruptions, signal-correlated noise, speech robotization, and unnatural male/female speech).
- Distortion classes and computation of the Mean Opinion Score (MOS) were designed for narrow-band telephone speech.
- Here, we propose new mappings based on subjective HA data from two different experiments.
- Internal features in P.563 were found to be highly correlated. In our models, features with a correlation coefficient higher than 0.9 with other features were discarded.
- Linear regression models were fitted using a stepwise procedure with the remaining features without considering feature interactions.

Experimental setup and results

Frequency lowering dataset [2]

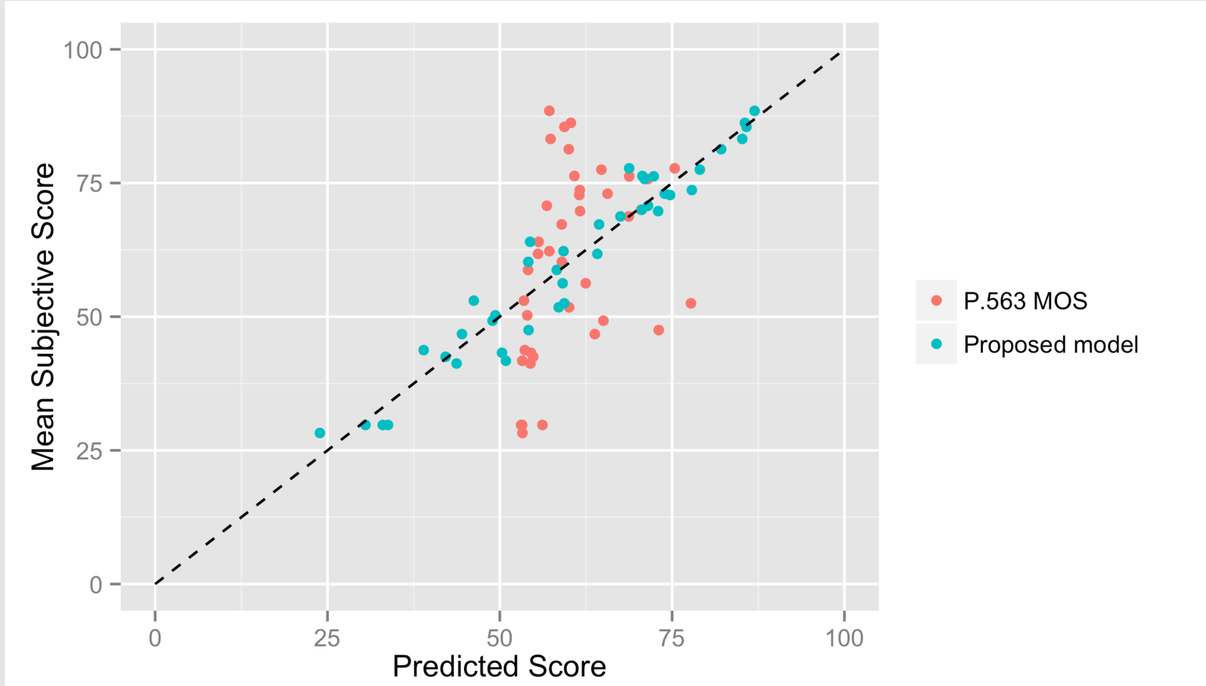
- Experiment explored the effect of frequency lowering (using nonlinear frequency compression) on perceived speech quality by HA users.
- Quality ratings for different talkers and NFC strategies (different cutoff frequencies and compression ratios) were obtained with 11 hearing impaired listeners with severe to profound hearing loss.
- A total of 32 different stimuli, with different NFC strategies (cutoff frequency/compression ratio) and anchors (low-pass filtering and clipping) were used.



(Intercept)	69.85 (0.59)***	fBasicVoiceQuality	32.35 (7.05)**
fSpeechLevel	-33.15 (9.02)*	fConsistentArtTracker	7.35 (4.94)
fPitchAverage	-56.10 (17.35)*	fVtpMaxTubeSection	-38.10 (13.40)*
fEstBGNoise	-45.20 (5.98)***	fFinalVtpAverage	35.75 (10.33)*
fSpecLevelDev	-71.69 (21.31)*	fVtpPeakTracker	37.67 (15.83)
fRelNoiseFloor	-12.29 (10.33)	fPitchCrossCorrelOffset	4.83 (4.24)
fSnr	8.40 (4.87)	fPitchCrossPower	32.10 (9.28)*
fHiFreqVar	238.38 (51.16)**	fUBeepsMean	-7.35 (1.96)**
fLocalBGNoiseLog	-23.13 (9.96)	fCepADev	-69.15 (14.46)**
fGlobalBGNoise	-16.83 (5.42)*	fCepSkew	-86.29 (17.57)**
fMuteLength	29.07 (5.03)*	fCepCurt	-108.28 (18.79)**
fSharpDeclines	-7.37 (4.62)	fLPCSkew	-26.82 (5.17)**
fUnnaturalSilenceMean	-5.41 (1.00)**	fPredictedMos	-38.06 (9.25)**
R	0.99	Num. obs.	32
Adj. R	0.96		

Speech enhancement dataset [3]

- Impact of HA speech enhancement on perceived quality was investigated in noise-only, reverberation-only, and noise-plus-reverberation listening conditions.
- Twenty two adult HA users with moderate to severe sensorineural hearing loss profiles rated degraded/enhanced stimuli using the MUSHRA quality scale.
- There were a total of 40 different conditions, combining two reverberation times (soundbooth - 0.1s and reverb chamber - 0.6 s), five noise types (quiet, stationary and babble at 0 and 5 dB), and four speech enhancement configurations (no enhancement, adaptive directional microphone, partial strength and full strength speech enhancement).



(Intercept)	60.46 (0.85)***	fUnnaturalSilenceMean	-11.55 (4.29)*
fEstSegSNR	7.85 (2.69)**	fBasicVoiceQuality	-30.81 (12.92)*
fSpecLevelDev	42.13 (13.84)**	fBasicVoiceQualityAsym	-30.53 (7.84)***
fSpecLevelRange	-21.25 (14.02)	fConsistentArtTracker	-14.00 (4.36)**
fRelNoiseFloor	-26.68 (7.83)**	fFinalVtpAverage	11.97 (3.69)**
fHiFreqVar	-19.68 (6.63)**	fCepCurt	41.63 (7.84)***
fLocalBGNoiseMean	-29.70 (7.59)***	fPredictedMos	-15.73 (7.85)
(Intercept)	60.46 (0.85)***	fUnnaturalSilenceMean	-11.55 (4.29)*
R	0.93	Num. obs.	40
Adj. R	0.90		

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

Conclusions

- Results show that while overall mean opinion score computed by P.563 has poor performance to predict enhanced speech quality for HA users, several internal parameters are highly correlated with the subjective ratings.
- We proposed mappings for two different speech processing strategies in HA, which showed a significant performance improvement when compared to a base model using only the P.563 MOS as a parameter.

References

[1] Rec, I. T. U. T. "P. 563: Single-ended method for objective speech quality assessment in narrow-band telephony applications." *International Telecommunication Union, Geneva* (2004).

[2] V. Parsa, S. Scollie, D. Glista, and A. Seelisch, "Nonlinear frequency compression effects on sound quality ratings of speech and music," *Trends in amplification*, vol. 17, no. 1, pp. 54–68, 2013.

[3] Keymanesh A, Scollie S, Parsa V, Folkeard P, Hayes D, Cornelisse L, and Allen P(2011). Perceptual Evaluation of Digital Signal Processing Strategies in a Modern Hearing Instrument across Noisy and Reverberant Environments. *Poster presented at the Canadian Academy of Audiology 14th Annual Conference*, Victoria.