IMPROVING QUALITY PREDICTION ACCURACY OF P.563 FOR NOISE SUPPRESSION

*L. Anders Ekman and W. Bastiaan Kleijn**

ACCESS Linnaeus Centre, Electrical Engineering
KTH – Royal Institute of Technology, 100 44 Stockholm, Sweden

ABSTRACT

We present a new method to map the features of P.563 to a single mean opinion score (MOS) value using non-negative matrix factorization (NMF). The method significantly improves the correlation performance for the case of speech databases containing noise suppression data, without affecting the performance for general speech databases.

Index Terms— Speech quality assessment, P.563, PESQ, NMF, noise suppression

1. INTRODUCTION

The perceived quality of speech is a highly subjective measure. To evaluate the subjective quality of speech processing systems, such as speech coders, noise suppressors, etc., time consuming and costly subjective listening tests are performed [1]. In a typical absolute category rating (ACR) test scenario [2], listeners (subjects) are asked to grade speech utterances on a scale from one to five, where one is bad and five is excellent.

The goal of objective quality measures is to mimic the behavior of the subjective listeners and algorithmically create an estimate of the speech quality. Objective quality assessment algorithms for speech are divided into intrusive and non-intrusive methods. The intrusive methods form the quality estimate using the clean test signal to compare to the processed, degraded signal. The non-intrusive methods use only the degraded signal to create the estimate. The current state-of-the-art methods in these two groups are the ITU-T standards P.862 (PESQ – perceptual evaluation of speech quality) [3] and P.563 [4], for intrusive and non-intrusive measurements, respectively. This paper focuses on the non-intrusive quality assessment methods, and in particular on P.563.

According to the standards, PESQ and P.563 are not validated for use with noise suppression algorithms [3, 4]. P.563 has demonstrated acceptable accuracy in transmission systems including echo cancellers and noise reduction systems under single talk conditions, but has not been validated for "effects and artifacts from isolated noise reduction algorithms" [4]. A recent study showed rather low performance of P.563 on noise suppression data [5]. Many current speech coders use noise suppression algorithms and, thus, it is crucial that the objective quality measures provides accurate estimates under such conditions.

A great need exists for a reliable non-intrusive quality assessment system for monitoring speech quality in telephone calls of live networks. For monitoring the quality of service (QoS) in live networks, the intrusive methods can not be used since the clean reference signal is not available [1]. Heterogenous networks, including voice communication over IP (VoIP) and wireless networks, create widely varying speech and noise environments that are difficult to predict. When different communication networks interact, using different technologies and equipment, complex distortions are introduced into the speech signal [6].

In this paper, we propose a simplification of the quality mapping function of P.563 that works well on speech processed with noise suppression algorithms and provides similar performance for other data. We replace P.563's complex mapping from speech features to objective score with an approach based on non-negative matrix factorization (NMF). NMF is well suited for finding additive structures in data, resulting in parts-based representations [7], often making intuitive interpretations possible.

The remainder of this paper is organized as follows. In section 2, we briefly describe the P.563 algorithm and discuss some of its shortcomings. In section 3 we describe the basics of NMF and our new application of it to non-intrusive quality assessment. Section 4 shows our experimental setup and our simulation results. Section 5 contains our conclusion.

2. P.563 FEATURES AND MAPPING

The P.563 algorithm [4] computes a large set of speech features that are mapped to an objective quality estimate. In total, there are 43 features, divided into five categories (number of features in parenthesis): i) mutes (4), ii) noise analysis (14), iii) unnatural speech (20), iv) basic speech descriptors (3) and v) speech extract parameters (2). The last group consists of outputs from a perceptual model that uses an intrusive perceptual speech quality measurement between the degraded signal and a pseudo reference signal found from enhancing the quality of the degraded signal.

The speech quality algorithm of P.563 consists of three main steps. In the first step, a distortion class is selected based on a set of features of the speech file under test. There are six different distortion classes: low static SNR, mutes and interruptions, low segmental SNR, unnatural voice – robotization, unnatural male voice and unnatural female voice. The distortion class is selected based on a logic scheme where, if the speech file is not considered belonging to class i, it is checked if it belongs to class $i + 1$, etc., see Figure 1.

The second step consists of computing a rough estimate of the speech quality based on that particular distortion class. In this step, different feature sets are used depending on the particular distortion class. For any given distortion class, a weighted sum of 12 features is calculated.

In the third step, a final quality estimate is computed as a weighted sum of the output of step two and an additional 11 features that are the same regardless of distortion class.

We have found that the logic scheme used to obtain the distortion class of a speech file makes P.563 less robust to changes and unseen distortion types. For any given test file, many features of speech are not used to form the quality estimate because of the distortion class selection. Furthermore, a small change in one of the features can

*This work was supported by Ericsson Research.

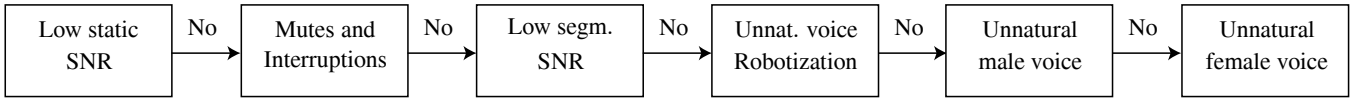

Fig. 1. Logic scheme of P.563 for selecting distortion class. In each box, a certain criterion is checked and if it applies, the speech file is classified as belonging to that box.

push the classifier into another class, which completely changes the mapping function and the resulting objective score. We made a small controlled experiment on 18 speech files that were on the boundary between being classified as "low static SNR" and "mutes and interruptions". The distortion is classified as the former if SNR$\leq$15 dB. We picked speech files that were within 0.1 dB of the 15 dB threshold and added or subtracted 0.1 dB from the estimated SNR to push it over to the other side of the threshold. From this small change in the SNR feature, the objective quality estimate changed by 0.36 MOS on average and the largest change was above 0.6. Thus, the quality estimate mapping of P.563 is very sensitive to small changes in individual features, which has a negative impact on robustness.

3. NON-NEGATIVE MATRIX FACTORIZATION

Non-negative matrix factorization is a method of approximate factorization of matrices with only non-negative entries. NMF (introduced as positive matrix factorization in [8]) introduces non-negativity constraints on all entries of the factorizing matrices. NMF has proven to be a powerful tool in a wide variety of applications, e.g., in finding the parts-based representation of images and in finding semantic features in text documents [7]. The non-negativity constraints result in basis vectors of additive structures in data, creating sparse representations that are interpretable when the input data has physical meaning.

In our application to quality assessment, we use NMF for finding the structures in the data that quantifies degradations in speech quality. The main competitor to NMF for finding structures in data for our application is principal component analysis (PCA). The use of NMF, when compared to PCA, gives a clear additive relationship for the degradation. In PCA, the constraint that the basis vectors must be orthogonal results in a representation of the original data that generally involves cancellations between positive and negative values. Also, the estimated objective score would consist of both adding and subtracting basis vectors, and thus the individual basis vectors would not represent a clear direction in terms of quality degradation. The non-negativity constraint of NMF leads to a clear interpretation of additive distortion structures in the feature space. Speech files for which the weighted sum of the basis vectors point in approximately the same direction also contain roughly the same type of perceived distortion.

3.1. NMF algorithm

Given a non-negative $n \times m$ matrix V that consists of m observation column vectors of dimension n, the non-negative matrix factorization results in the approximation

$$V \approx WH, \tag{1}$$

where the factorizing matrices W and H also are non-negative. The $n \times r$ matrix W consists of r basis vectors and the $r \times m$ matrix H is the data represented in the new basis. The quality of the factorization approximation of (1) is determined with a cost function. Lee and Seung [9] use two different cost functions: the square of the Euclidian distance and a measure similar to the Kullback-Leibler divergence. In this paper, we use the square of the Euclidian distance, $||V - WH||^2$, since we found that this measure gives better results for our application. For this distortion criterion, [9] provides multiplicative update rules that guarantee a monotonic convergence towards a local optimum.

Initializing the W and H matrices using PCA [10] has proven to give good results for our application and is therefore used in this paper. For initializing the NMF matrices, PCA projects the data onto a lower dimensional linear space (to match the dimension r of the NMF matrices), such that the variance of the projected data is maximized. This means that the PCA finds the directions of highest energy in the data space, which has proven to be a good starting point for the NMF. In our experiments, we have found that the results from using PCA initialization are similar to the results of the best NMF solutions from a large set of randomized initializations.

We have found that our application benefits from sparseness constraints. Hoyer [11] presents a method to explicitly control the degree of sparseness in the solution by defining a sparseness measure and incorporating it into the NMF algorithm. The sparseness measure $\mathcal{S}$ is defined on the columns of W and on the rows of H by

$$\mathcal{S}(w_i) = \frac{1}{\sqrt{n}-1}\left(\sqrt{n} - \frac{\sum_{i=1}^{n}|w_{ij}|}{\sqrt{\sum_{i=1}^{n} w_{ij}^2}}\right) = \mathcal{S}_w, \quad \forall i \tag{2}$$

$$\mathcal{S}(h_j) = \frac{1}{\sqrt{m}-1}\left(\sqrt{m} - \frac{\sum_{j=1}^{m}|h_{ij}|}{\sqrt{\sum_{j=1}^{n} h_{ij}^2}}\right) = \mathcal{S}_h, \quad \forall j. \tag{3}$$

The measure evaluates to unity for vectors that contain only a single non-zero component, and takes the value zero for vectors where all components are equal. A sparse W means that the basis vectors are sparse, so that few structural properties of the data are captured in each basis vector. A sparse H indicates that each data vector is reconstructed from only a few basis vectors.

3.2. NMF applied to features of P.563

We propose a system for quality assessment using NMF with the speech features of P.563. The speech features and subjective score of each utterance are grouped together to form the non-negative $n \times m$ matrix V, which corresponds to the feature vectors versus the observation index. The first row of V is selected to be $5 - Q$ for the database, where Q denotes the subjective MOS. The first row of W corresponds to degradation in MOS, and the basis vectors scale with quality degradation. The addition of basis vectors to form a certain feature vector corresponds to adding distortions together, and the first entry of the basis vectors indicate how severe the impact of that particular basis vector is on the quality. Basis vectors with high values in the first entry represent structures that have a strong negative impact on speech quality. The basis vectors with low values in the first entry have low impact on the speech quality. These vectors characterize structure that does not affect speech quality.

To create a quality estimate for a new speech utterance, we must rely on the feature vector without its first entry, since the subjective

score is not known. Let $\widetilde{v}$ denote the feature vector (dimension $n-1$) of the new utterance. We need to infer the hidden variables in h so that

$$\widetilde{v} \approx \widetilde{W} h, \tag{4}$$

where the $(n-1)\times r$ matrix $\widetilde{W}$ is the lower $n-1$ rows of W. We find the h that minimizes $||\widetilde{v}-\widetilde{W}h||^2$ under the non-negativity constraint $h \geq 0$ using the method of Lagrange multipliers. Given the hidden variable vector h, which is full-size ($r \times 1$), the best approximation of the complete feature vector $\hat{v}$ is

$$\hat{v} = Wh, \tag{5}$$

and the first entry of $\hat{v}$ gives us the quality estimate for the utterance.

A large majority of the features of P.563 are non-negative. For the features that have negative entries, we have made small adjustments so that they fulfill the non-negativity constraint. Furthermore, the feature values f are linearly mapped from $[\min(f), \max(f)]$ to the unit interval $[0, 1]$, so that each feature receives roughly the same importance in the cost criterion of the NMF.

4. SIMULATION

We extracted the features of P.563 for a set of databases as training data, comprising the data matrix V in (1). These training databases are seven databases from P Supplement 23 [12], experiments 1 and 3, consisting of speech processed with speech coders G.711, G.726, G.728, G.729, GSM-FR, IS-54 with and without bit errors and frame erasures and with and without background noise. The P Supplement 23 databases composed parts of the training data used in the development of P.563 [6]. Our training data further consists of two databases containing speech in background noise (car and babble noise) processed with the AMR-NB codec and three different noise suppression algorithms developed at Ericsson. The training data set consists of in total 388 conditions and 2672 processed speech files. For our NMF mapping, we used the parameters $r = 32$, $\mathcal{S}_w = 0.38$, and a relaxed sparseness constraint on H, as described in section 3. We applied the NMF algorithm and the resulting basis vectors W formed our trained quality assessment model.

The performance of the objective quality assessment methods is measured by the correlation coefficient R (also known as Pearson's correlation coefficient) and the root mean squared error (RMSE) ε,

$$R = \frac{\sum_i (Q_i - \mu_Q)(\hat{Q}_i - \mu_{\hat{Q}})}{\sqrt{\sum_i (Q_i - \mu_Q)^2 \sum_i (\hat{Q}_i - \mu_{\hat{Q}})^2}} \tag{6}$$

$$\varepsilon = \sqrt{\frac{1}{N}\sum_i (Q_i - \hat{Q}_i)^2}, \tag{7}$$

where Q_i denotes the individual subjective scores, μ_Q is the average subjective score, N is the number of samples, and corresponding variables for the objective scores are denoted with $\hat{Q}$. The correlation coefficient and RMSE are measured on a per-condition basis. The subjective and objective scores are first averaged for each test condition (e.g., type of speech codec, data rate, degree of noise, etc.) of the speech database, and then the correlation coefficient and RMSE are computed. To account for differences between databases and the fact that identical distortions can give rise to different subjective quality scores in different studies, we follow the standard procedure to use a monotonic third order polynomial mapping from objective score onto the subjective score for each individual database [3, 4].

	NMF		P.563		PESQ	
Database	R	ε	R	ε	R	ε
NS 1	0.89	0.33	0.71	0.51	0.98	0.15
NS 2	0.95	0.23	0.71	0.51	0.94	0.25
NS 3	0.93	0.27	0.65	0.56	0.93	0.28
Average	**0.92**	0.28	**0.69**	0.53	**0.95**	0.23

Table 1. **Validation data.** Correlation coefficients and RMSE for P.563 and for our proposed NMF mapping with the features of P.563. The performance of PESQ is also shown as reference. The values are after averaging over conditions and using a monotonic third order polynomial fit from subjective to objective score.

To validate the performance, we used three unseen databases as validation set. The validation data set consists of in total 78 conditions in 1560 processed speech files, comprising of speech degraded by car noise (NS 1), street noise (NS 2) and babble noise (NS 3) processed by the AMR-NB codec with and without noise suppression algorithms. The approach described by (4)-(5) was applied to create the quality estimates for all speech files in our validation data set.

Figure 2 shows a scatter plot of all conditions of the validation database set. The mapping of P.563 has problems with the quality estimates for these databases. The correlation coefficient and RMSE for the P.563 mapping and using our NMF mapping are shown in Table 1. It is seen that our mapping outperforms the original mapping of P.563 for these noise suppression databases. The simple yet powerful mapping of NMF finds the structures in the feature space that correspond to degradations in quality and quantifies them, giving a certain degree of degradation. The sparseness constraint parameter choice of $S_w = 0.38$ gives the best result in terms of average correlation coefficient for the NS databases, but parameter values in the range 0.35–0.45 work well, with average correlation coefficients always above 0.88. We note the fact that P.563 was not trained using noise suppression data, and so it has an inherent disadvantage over our proposed method in these experiments. We do not claim that our method necessarily would outperform P.563 if also P.563 were to be trained on databases containing noise suppression data but the NMF approach provides a more intuitive and simpler structure.

The sensitivity of our method for small changes in a single feature was investigated through the same experiment as described in the final paragraph of section 2. A small change in the SNR feature resulted in an average change of 0.14 in the objective quality estimate (compared to 0.36 for P.563), indicating that our method is more robust to small changes in individual features. Our method also works well in the case of general speech and distortions, shown by Table 2. The table shows the performance of our NMF approach and P.563 over the part of the training database that does not contain noise suppression data. The NMF mapping approach shows equivalent performance as P.563 on the P supplement 23 training databases: $R_{\mathrm{NMF}} = 0.87$ versus $R_{\mathrm{P.563}} = 0.88$ averaged over the seven databases. Note that these databases also were used in the training of P.563 [6].

Table 1 shows also the results from PESQ as reference. PESQ performs well for our noise suppression databases, and similar trends have been reported previously, e.g., in [13].

5. CONCLUSION

We have presented a new method to create a mapping from the features of P.563 to a single objective quality score using non-negative matrix factorization. The power of NMF lies in its ability to extract

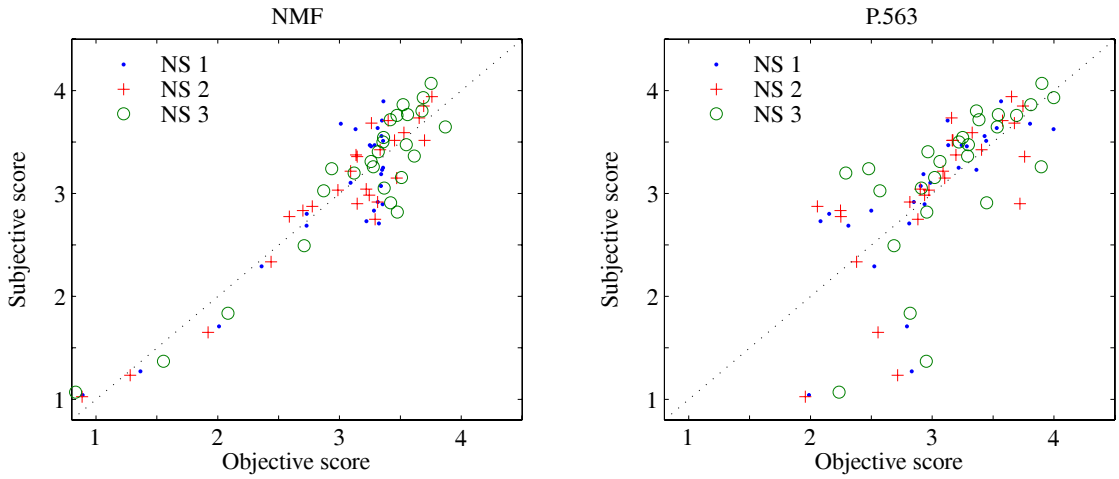

Fig. 2. Scatter plot of the subjective versus objective score for the three validation databases. The values are averaged per condition and a third order monotonic polynomial was applied. The three different markers indicate the three different databases in the validation data set.

	NMF		P.563		PESQ	
Database	R	ε	R	ε	R	ε
P 23 exp 1A	0.89	0.34	0.89	0.34	0.94	0.25
P 23 exp 1D	0.88	0.30	0.80	0.38	0.96	0.18
P 23 exp 1O	0.91	0.32	0.92	0.31	0.96	0.20
P 23 exp 3A	0.85	0.36	0.87	0.34	0.90	0.29
P 23 exp 3C	0.82	0.47	0.85	0.44	0.97	0.22
P 23 exp 3D	0.89	0.30	0.93	0.25	0.95	0.22
P 23 exp 3O	0.85	0.38	0.91	0.30	0.93	0.26
Average	**0.87**	0.35	**0.88**	0.34	**0.95**	0.23

Table 2. **Training data.** Correlation coefficients and RMSE for P.563 and for our proposed NMF mapping on the P Supplement 23 database set used for training. The performance of PESQ is also shown as reference. The P Supplement 23 databases contain speech in French (A), Italian (C), Japanese (D) and American English (O). The values are after averaging over conditions and using a monotonic third order polynomial fit from subjective to objective score.

the larger features from speech that correspond to a certain type of degradation in quality, and to grade how severe the degradation is. Our method has a straightforward algorithmic structure, and it outperforms the mapping of P.563 in the databases containing speech processed with noise suppression systems. It also shows equivalent performance for general speech data. This indicates that the mapping of P.563 is the weak part of the standard, rather than the selected set of features.

6. ACKNOWLEDGEMENT

The authors would like to thank Volodya Grancharov of Ericsson Research for providing us with valuable ideas for this work, as well as the speech databases.

7. REFERENCES

[1] Volodya Grancharov and W. Bastiaan Kleijn, "Speech quality assessment," in *Springer Handbook of Speech Processing*, Jacob Benesty, M. M. Sondhi, and Yiteng Huang, Eds., pp. 83–102. Springer, 2007.

[2] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," 1996.

[3] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ)," 2001.

[4] ITU-T Rec. P.563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," May 2004.

[5] Tiago H. Falk, Hua Yuan, and Wai-Yip Chan, "Single-ended quality measurement of noise suppressed speech based on Kullback-Leibler distances," *Journal of Multimedia*, vol. 2, no. 5, pp. 19–26, September 2007.

[6] L. Malfait, J. Berger, and M. Kastner, "P.563 - the ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 1924–1934, Nov. 2006.

[7] Daniel D. Lee and H. Sebastian Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[8] Pentti Paatero, Unto Tapper, Pasi Aalto, and Markku Kulmala, "Matrix factorization methods for analyzing diffusion battery data," *Journal of Aerosol Science*, vol. 22, pp. S273–S276, 1991.

[9] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2000, vol. 13, pp. 556–562.

[10] Zhonglong Zheng, Jie Yang, and Yitan Zhu, "Initialization enhancer for non-negative matrix factorization," *Engineering Applications of Artificial Intelligence*, vol. 20, pp. 101–110, 2007.

[11] Patrik O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[12] ITU-T Rec. P. Supplement 23, "ITU-T coded-speech database," 1998.

[13] Thomas Rohdenburg, Volker Hohmann, and Birger Kollmeier, "Objective perceptual quality measures for the evaluation of noise reduction schemes," in *Proc. 9th Intl. Workshop on Acoustic Echo and Noise Control*, 2005, pp. 169–172.