Speech Enhancement Quality Assessment Based on Aspect-Specific Qualities: A Preliminary Analysis

1st Benjamin Stahl
Institute of Electronic Music and Acoustics
University of Music and Performing Arts
Graz, Austria
stahl@iem.at

2nd Alois Sontacchi
Institute of Electronic Music and Acoustics
University of Music and Performing Arts
Graz, Austria
sontacchi@iem.at

***Abstract*—We propose a methodical framework to develop computational tools for assessing the quality of enhanced speech signals. The central building block of this framework are aspect-specific subjective quality ratings obtained in a listening experiment. We show that mean aspect-specific subjective quality ratings predict overall quality significantly better than objective features of state-of-the-art quality assessment tools. These experimentally obtained aspect-specific subjective quality features can be utilized to determine and tune objective features that predict them and thus indirectly predict overall quality.**

***Index Terms*—speech quality, speech enhancement, computational quality assessment**

I. Introduction

Defined listening test procedures and computational models for subjectively and objectively assessing the perceived quality of speech enhancement algorithms have first been necessitated by the standardization of speech transmission systems [1], [2]. More recently, quality prediction has gained attention, when campaigns for the evaluation of source separation algorithms were introduced [3], [4].

Some quality prediction approaches compute quality as a single similarity/distance between time-frequency (TF) representations of a *test stimulus* and a *reference stimulus*. The frequency-weighted segmental signal-to-noise ratio (fwsegSNR) [5, ch. 11] combines perceptual and information-theoretical principles by computing the short-time SNR in frequency bands that are proportional to the ear's critical bands, and by computing a weighted average of these values. Another approach, PEMO-Q [6], computes the similarity metric PSM_t between biomimetic TF representations of the reference and the test stimulus and maps it onto overall quality using a piecewise rational function. Similarly, the speech variant of ViSQOL [7], [8] computes a *Neurogram Similarity Index Measure* (NSIM) motivated by the Structural Similarity Index (SSIM), which is used in image quality assessment. The NSIM measure is then mapped onto overall quality using a polynomial function.

Other computational assessment methods model quality as a function of multiple features. The Perceptual Evaluation of Speech Quality (PESQ) [2] computes an aggregated absolute difference D between time- and gain-aligned TF loudness representations of the reference and the test stimulus. Additionally, it computes the asymmetric difference D^{asym} that only takes into account TF components in which the power density of the test stimulus is clearly greater than the power density of the reference stimulus. Such components more likely lead to separate auditory objects and thus are perceived as more disturbing. The overall quality predicted by PESQ is a sigmoid-like mapping of a linear combination of D and D^{asym}. The Perceptual Evaluation of Audio Source Separation (PEASS) [4], [9] employs a more sophisticated approach. Based on the true audio source signals (target source and interference sources), it decomposes the difference between reference and test stimulus into three components: target distortion, interference, and artifacts. It then computes the perceptual similarities PEMO-Q PSM_t between different *less corrupted* versions of the test stimulus (by removing different corruption components) and the test stimulus. These values are then mapped onto overall quality and onto three aspect-specific qualities using simple neural networks. The audio variant of ViSQOL [8] uses per-frequency-band NSIMs and maps them onto overall quality using a Support Vector Regression. The Perceptual Evaluation of Audio Quality (PEAQ) [10] computes 11 psycho-acoustically motivated features, called model output variables (MOVs), and maps them onto overall quality using a neural network. Kastner's 2f-model [11] uses only two of these MOVs, which are mapped onto overall quality by a rational function.

Subjective audio quality can be described as "the perceptual distance between a set of [namable] sound-character features and a set of reference features" [12], [13, ch. 3]. Here, we refer to the distances between individual (sets of) sound-character features as *aspect-specific qualities*. Examples for such aspect-specific qualities are disturbance by background sounds and preservation of the target signal, which could be divided into sub-aspects such as equality of timbre or equality of amount of reverberation. The computational quality assessment methods discussed here compute overall quality as a direct function of *objective features*, i.e., D and D^{asym} with PESQ or the per-frequency-band NSIMs with ViSQOL. While some methods' objective features aim to capture individual subjective aspect-specific qualities, the objective features are typically not

The experiment participant remuneration was funded by the *Forschungsinstitut für Elektronische Musik und Akustik (FiEMA)*, Graz, Austria.

individually tuned to map on these aspect-specific qualities in a one-to-one way. Subjective aspect-specific ratings are most often not available. Instead, objective features are usually tuned according to the developers' domain knowledge and judgment and then mapped onto subjective overall quality.

In this publication, we propose an alternative methodical framework for the computational assessment of speech enhancement quality. Figure 1 shows a schematic representation of this framework. Specifically tuned objective features are used to model subjective aspect-specific qualities (obtained in listening experiments), which in turn are used to model overall quality. Such a framework — with namable subjective aspect-specific qualities — closely reproduces the formation of an overall quality concept in human perception [13, ch. 3]. Therefore, using formally collected subjective aspect-specific ratings to select and tune objective features and modeling the two stages of quality formation shown in Fig. 1 is a promising approach to computational assessment of overall perceived quality.

The proposed framework is novel insofar as that with PEASS, aspect-specific quality ratings (target preservation, presence of other sources, presence of artificial noise) were indeed collected in addition to overall quality ratings in a listening experiment, but overall quality predictions are not computed from predictions on these aspect-specific qualities, but rather directly from the same objective features that are used to predict aspect-specific qualities. The aspect-specific qualities proposed with PEASS are a possible set of aspect-specific qualities to be used within the proposed framework. A different set of aspect-specific qualities is proposed by the ITU recommendation for *subjective* quality assessment of speech enhancement [1]. It proposes a listening experiment in which participants first rate a stimulus' target preservation, then the disturbance of background sounds, and finally the overall quality by subjectively weighting the two aspects.

This publication should be considered a preliminary verification of the feasibility of the proposed framework. Only if experimentally obtained aspect-specific quality ratings predict overall quality clearly better than the objective features of state-of-the-art computational quality assessment methods, such a framework can lead to an improvement in computational quality assessment. Thus, we formulate the following research question: *Do subjective aspect-specific quality ratings predict overall quality better than the best state-of-the-art objective features?*

II. Methods

A. Collection of subjective ratings

We collected subjective ratings of aspect-specific qualities and overall quality of enhanced speech stimuli in an experiment. Twenty-six participants, aged 20 to 58, took part in the experiment. All participants work or study in the field of audio. The web-based listening experiment software webMUSHRA [14] was used. Participants used full-size headphones.

Stimuli were created from different noisy speech mixtures and speech enhancement algorithms. The stimuli and subjective ratings used in [4] are publicly available. This dataset includes 20 real source separation output stimuli from five different

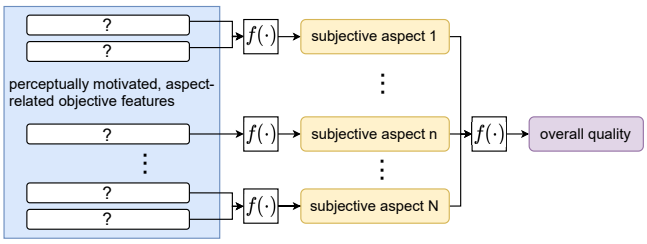

Fig. 1. Exemplary model in the proposed framework.

mixtures in which speech is the target signal. However, more datapoints are required for the multiple local regression that we applied to determine the overall quality prediction strength of different feature sets. Therefore, we added 11 more mixtures, each also processed with four speech enhancement algorithms, containing different interference signals and target speech signals. Six of these mixtures were created using semi-anechoic source signals and a room impulse response simulation with two virtual microphones and moving speakers. The speaker signals in these mixtures were taken from the EBU SQAM material [15], the interference source signals were typical office sounds. The other five mixtures were simulated mixtures from the CHiME3 [16] challenge. Each of these mixture signals was processed with four speech enhancement methods. The employed speech enhancement algorithms include both traditional and neural network mask based beamforming and postfilter approaches [17]–[19]. In total, 16 mixtures (=64 stimuli) were thus available for rating. With each mixture, the clean target speech signal is considered the reference stimulus. All stimuli have a duration of 5 s, are sampled at $f_s = 16\,\mathrm{kHz}$, and are loudness-matched following [20]. An anchor stimulus was created for each mixture by adding a mix of all interference signals and an artifacts signal to a corrupted version of the target signal. This stimulus combines the signal degradations of the different anchor stimuli used in [4].

Fifteen mixtures were divided into three mixture groups, each containing five mixtures. The 26 participants were distributed across the mixture groups, so that each participant only rated five mixtures. The remaining mixture was used as a training mixture. Each participant successively performed the following tasks on the assigned stimuli: *Rate the …*

- *… test sounds with respect to the* ***overall quality****.*
- *… quality of the test sounds with respect to the* ***target preservation****. For this purpose, only focus on the target (i.e., the sound associated with the reference) in the test sounds and disregard all other sounds.*
- *… quality of the test sounds with respect to the* ***absence of additional ARTIFICIAL sounds****.*
- *… quality of the test sounds with respect to the* ***absence of other NATURAL sound sources****.*
- *… quality of the test sounds with respect to the* ***disturbance of background sounds*** *(****ALL*** *artificial and natural sounds, except for the sound associated with the reference).*

The overall quality rating was always performed first, while the other four tasks were performed in randomized order. This

guarantees that the aspect-specific tasks do not influence the participants' criteria for overall quality and that no bias emerges from the order of rating tasks.

The Multi-Stimulus with Hidden Reference and Anchor (MUSHRA) [21] method was used for the rating. For each of the tasks described above, the participants first did a training page with the training mixture and then five test pages, where on each page they jointly rated the four enhanced signals, a hidden reference signal, and an anchor signal associated with one mixture. The order of pages/mixtures as well as the arrangement of stimuli on each page was randomized. The function of the hidden reference and anchor signals is to spread the quality range on each MUSHRA page and decrease context effects in this way, so ratings from different MUSHRA pages stay comparable. The ratings of these signals, however, were only used to assess the soundness of ratings. The data of one participant had to be removed, since the participant consistently did not rate the hidden reference with the maximum score.

The obtained ratings of overall quality, target preservation, presence of artificial sounds, presence of other natural sources, and overall disturbance of background sounds are denoted by $y_{p,m,s}$, $a^{\text{target}}_{p,m,s}$, $a^{\text{artif}}_{p,m,s}$, $a^{\text{other}}_{p,m,s}$, and $a^{\text{backg}}_{p,m,s}$. The subscripts p, m, and s respectively denote the participant, the mixture scenario, and the stimulus in each mixture (output signal of one of four speech enhancement algorithms).

B. Computation of objective features

We computed the objective features and quality predictions of all computational quality assessments described in Section I using Loizou's MATLAB scripts [5], PEASS v2.0.1 [9] (and implicitly the low-pass version of PEMO-Q [6]), ViSQOL v3 [8], the PQEvalAudio implementation of PEAQ [22], and the formulae given in [6] and [23].

C. Determining prediction strength of feature sets

1) Cross-validation procedure: By averaging the participants' overall ratings over a subset $\mathcal{P}_\subseteq$ of participants, the mean overall quality rating $\bar{y}_{\mathcal{P}_\subseteq,m,s}$ can be obtained for each stimulus (m,s). Note that due to the between-participants division of mixtures, mean overall ratings can only be obtained from those participants in $\mathcal{P}_\subseteq$ that rated the respective stimuli. Let $\overline{\mathbf{x}}^{\xi}_{\mathcal{P}_\subseteq,m,s}$ be a vector of *features*, which can be across-participant mean ratings on different aspect-specific qualities obtained from the rating tasks described in Section II-A or different objective features obtained from the algorithms named in Section II-B. The superscript ξ denotes the choice of feature set. Note that objective features are independent of the subscript $\mathcal{P}_\subseteq$, but included in this definition for unification. The regression model $f(\overline{\mathbf{x}},\mathbf{h},\mathbf{c})$ with model hyperparameters $\mathbf{h}$ and model coefficients $\mathbf{c}$ is applied to map the features $\overline{\mathbf{x}}^{\xi}_{\mathcal{P}_\subseteq,m,s}$ onto an estimate of the mean overall quality rating. The optimal (with respect to some loss function) coefficients of a model fitted to features and overall quality ratings from the subset of participants $\mathcal{P}_\subseteq$ on the subset of mixtures $\mathcal{M}_\subseteq$ are denoted by $\mathbf{c}_{\mathcal{P}_\subseteq,\mathcal{M}_\subseteq}$. To ascertain the prediction strength of models with different feature vectors and hyperparameters with unseen data, we propose the following cross-validation procedure: Let $\mathcal{P}_{\setminus e}$ and $\mathcal{M}_{\setminus e}$ be the subsets of participants and mixtures, that each exclude one participant or one mixture. The model $f\left(\overline{\mathbf{x}},\mathbf{h},\mathbf{c}_{\mathcal{P}_{\setminus e},\mathcal{M}_{\setminus e}}\right)$ fitted to the mean ratings of participants in $\mathcal{P}_{\setminus e}$ on the stimuli in $\mathcal{M}_{\setminus e}$ is then used to predict the mean overall ratings $\bar{y}_{\mathcal{P},m_e,s}$ on the stimuli of the excluded mixture m_e from *all* participants rating that mixture using $\overline{\mathbf{x}}^{\xi}_{\mathcal{P},m_e,s}$, the mean aspect-specific ratings / objective features associated with the stimuli in m_e:

$$\hat{\bar{y}}_{\mathcal{P},m_e,s\mathcal{P}_{\setminus e}}(\xi,\mathbf{h}) = f\left(\overline{\mathbf{x}}^{\xi}_{\mathcal{P},m_e,s},\mathbf{h},\mathbf{c}_{\mathcal{P}_{\setminus e},\mathcal{M}_{\setminus e}}\right). \tag{1}$$

The mean squared error is then computed as

$$\mathrm{MSE}_{\mathcal{P}_{\setminus e},m_e}(\xi,\mathbf{h}) = \frac{1}{\#\mathcal{S}}\sum_{s\in\mathcal{S}}\left(\hat{\bar{y}}_{\mathcal{P},m_e,s\mathcal{P}_{\setminus e}}(\xi,\mathbf{h}) - \bar{y}_{\mathcal{P},m_e,s}\right)^2, \tag{2}$$

where $\#$ denotes the number of elements in a set. The expected value of $\mathrm{MSE}_{\mathcal{P}_{\setminus e},m_e}(\xi,\mathbf{h})$ is the expected squared error when predicting the mean rating of an *unknown* stimulus with a *limited* sample of ratings on other stimuli. Leaving out one participant when fitting the model accounts for the uncertainty of mean subjective ratings. The cross-validation is repeated for all possible combinations of excluded participants and mixtures. Since $\mathrm{MSE}_{\mathcal{P}_{\setminus e},m_e}(\xi,\mathbf{h})$ cannot be considered independent across subsets $\mathcal{P}_{\setminus e}$ within the mixture m_e, the mixture means $\overline{\mathrm{MSE}}_{m_e}(\xi,\mathbf{h})$ must be considered for comparative inference on feature sets and hyperparameters.

2) Regression models: As a realization of $f(\overline{\mathbf{x}},\mathbf{h},\mathbf{c})$, we used k-nearest neighbor locally weighted regression [24]. The first hyperparameter, r, defines the fraction of datapoints used for the local regression, such that $k = \lfloor r\cdot D\rfloor$, where D is the number of total datapoints and k is the number of used datapoints. A Gaussian weighting w is defined as

$$w(\overline{\mathbf{x}},\overline{\mathbf{x}}_0) = \exp\left(-\frac{d^2(\overline{\mathbf{x}},\overline{\mathbf{x}}_0)}{d^2(\overline{\mathbf{x}}_{k+1},\overline{\mathbf{x}}_0)\cdot 2s^2}\right), \tag{3}$$

where $\overline{\mathbf{x}}_0$ is the point at which $f(\overline{\mathbf{x}},\mathbf{h},\mathbf{c})$ is evaluated, $d(\overline{\mathbf{x}},\overline{\mathbf{x}}_0)$ is the Euclidean distance between $\overline{\mathbf{x}}$ and $\overline{\mathbf{x}}_0$, and $\overline{\mathbf{x}}_{k+1}$ is the *(k+1)*-th nearest point to $\overline{\mathbf{x}}_0$. The parameter s is a scale factor, the second hyperparameter we define for our regression model. As a third hyperparameter, we use the polynomial degree p of the local regression. The function $f(\overline{\mathbf{x}},\mathbf{h},\mathbf{c})$ at $\overline{\mathbf{x}}_0$ is defined as weighted degree-p polynomial regression with the k nearest neighbor datapoints $\overline{\mathbf{x}}_n, n\in\{1,\ldots,k\}$, using $w(\overline{\mathbf{x}}_n,\overline{\mathbf{x}}_0)$ as weights. Thus, the found regression coefficients are location-dependent. Note that the regression models include standardization of the regressors. The hyperparameters control the following trade-offs:

- For small r and large p, the model tends to overfit; for large r and small p, it tends to underfit.
- For small s, more weight is given to closer datapoints, leading to less discontinuities at regions at which the set of k nearest neighbors change; for large s, the k datapoints are given more equal weights.

At $r=1$ and $s=\infty$, the regression is a global regression.

3) Evaluated feature sets: We considered the following (vectors of) *objective features* as realizations of the feature vector $\overline{\mathbf{x}}^{\xi}_{\mathcal{P}_{\subseteq},m,s}$ (note that we omit the subscripts $\mathcal{P}_{\subseteq}$, m, and s for readability, when referring to features):

- fwsegSNR
- NSIM
- $[D^{\mathrm{PESQ}}, D^{\mathrm{PESQ}}_{\mathrm{asym}}]$ and the two features it is comprised of
- the PEASS feature vector $[\mathrm{PSM}_t, \mathrm{PSM}_t^{\mathrm{target}}, \mathrm{PSM}_t^{\mathrm{interf}}, \mathrm{PSM}_t^{\mathrm{artif}}]$ and the 14 vectors resulting from combining subsets of its elements (including single features)
- $[\mathrm{AvgModDiff1}, \mathrm{ADB}]$ and its two single features
- the full set of ViSQOL-audio features
- the full set of PEAQ features

As can be seen, selecting features by fitting models on subsets of feature sets was not done with the features of ViSQOL-audio and PEAQ. Due to the large number of possible combinations, this was computationally not feasible. Next to the objective features, we considered the following (vectors of) *aspect-specific subjective ratings*:

- subjective superset 1 (associated with PEASS): $[\bar{a}^{\mathrm{target}}, \bar{a}^{\mathrm{artif}}, \bar{a}^{\mathrm{other}}]$ and the six vectors resulting from combining subsets of its elements
- subjective superset 2 (associated with ITU-T P.835): $[\bar{a}^{\mathrm{target}}, \bar{a}^{\mathrm{backg}}]$ and the single feature $\bar{a}^{\mathrm{backg}}$

To find the optimal hyperparameters for each feature vector, we varied r from 0.1 to 1 in steps of 0.1 and s from $10^{-0.5}$ to $10^{0.5}$ in 10 logarithmically spaced steps. We appended $s = \infty$ to the evaluated values of s. We fitted weighted mean, linear, and quadratic regression models ($p \in \{0, 1, 2\}$). In order to also evaluate the fitted mappings of the discussed objective quality assessment methods themselves, we additionally considered the *overall quality estimates* of these methods as realizations of $\overline{\mathbf{x}}^{\xi}_{\mathcal{P}_{\subseteq},m,s}$. However, we only used global linear regression models ($r = 1, s = \infty, p = 1$) with these overall quality estimates, since they are already the output of nonlinear mappings and should be proportional to mean subjective quality. The hyperparameters associated with the lowest mean $\overline{\mathrm{MSE}}_{m_e}(\xi, \mathbf{h})$ are denoted by $\mathbf{h}_{\mathrm{opt}_\xi}$. The cross-validation mean squared errors $\overline{\mathrm{MSE}}_{m_e}(\xi, \mathbf{h}_{\mathrm{opt}_\xi})$ reflect the prediction strength of the feature set ξ. From here on, we omit the hyperparameters $\mathbf{h}_{\mathrm{opt}_\xi}$ for readability: $\overline{\mathrm{MSE}}_{m_e}(\xi) \coloneqq \overline{\mathrm{MSE}}_{m_e}(\xi, \mathbf{h}_{\mathrm{opt}_\xi})$.

4) Comparisons between feature sets: We computed the mean of $\overline{\mathrm{MSE}}_{m_e}(\xi)$ for each feature set ξ. To visualize confidence in the central tendencies of the distributions, we also computed the median and its binomial-distribution-based 95% confidence interval. The distribution of $\overline{\mathrm{MSE}}_{m_e}(\xi)$ is non-Gaussian. Therefore, we carried out pairwise comparisons of different feature sets ξ using Wilcoxon signed-rank tests [25]. To test our main hypothesis, we selected the respective aspect-specific subjective feature sets associated with the lowest mean of $\overline{\mathrm{MSE}}_{m_e}$ out of the two supersets of aspect-specific subjective features enumerated in Section II-C3. Furthermore, we selected the objective feature set associated with the lowest mean of $\overline{\mathrm{MSE}}_{m_e}$. To compare each of the two best subjective aspect-specific feature sets to the best objective feature sets, we then carried out two one-sided Wilcoxon signed-rank tests using the respective $\overline{\mathrm{MSE}}_{m_e}$ values ($N = 15$ mixtures). We set the global significance level to $\alpha = 0.05$ and jointly tested the two hypotheses in a Bonferroni-Holm procedure [26].

Our stimuli, listening experiment configuration, subjective results, and MATLAB analysis scripts are available online[1].

III. Result

Figure 2 visualizes $\overline{\mathrm{MSE}}_{m_e}(\xi)$ associated with objective feature sets (black), objective quality predictions (violet), and feature sets of across-participant mean aspect-specific quality ratings (dark yellow). Note that some objective features are shown multiple times, since they are used in multiple objective quality assessment methods. We identified $[\bar{a}^{\mathrm{target}}, \bar{a}^{\mathrm{other}}]$ as the best feature set out of subjective superset 1, $[\bar{a}^{\mathrm{target}}, \bar{a}^{\mathrm{backg}}]$ as the best feature set out of subjective superset 2, and $[\mathrm{PSM}_t, \mathrm{PSM}_t^{\mathrm{interf}}, \mathrm{PSM}_t^{\mathrm{artif}}]$ as the best objective feature set. We compared the $\overline{\mathrm{MSE}}_{m_e}$ values associated with $[\bar{a}^{\mathrm{target}}, \bar{a}^{\mathrm{other}}]$ to those associated with $[\mathrm{PSM}_t, \mathrm{PSM}_t^{\mathrm{interf}}, \mathrm{PSM}_t^{\mathrm{artif}}]$ in a one-sided Wilcoxon signed-rank test [25] and obtained $\boldsymbol{p_1 = 0.0177}$. Comparing the $\overline{\mathrm{MSE}}_{m_e}$ values associated with $[\bar{a}^{\mathrm{target}}, \bar{a}^{\mathrm{backg}}]$ and $[\mathrm{PSM}_t, \mathrm{PSM}_t^{\mathrm{interf}}, \mathrm{PSM}_t^{\mathrm{artif}}]$ yielded $\boldsymbol{p_2 = 0.0240}$. Both differences are significant at their respective Bonferroni-Holm-corrected [26] significance levels ($\alpha_1 = 0.025, \alpha_2 = 0.05$).

IV. Discussion

We found that the respective best subjective feature sets out of the supersets associated with PEASS and ITU-T P.835 both predict overall quality significantly better than the best objective feature set found. Interestingly, out of the subjective superset associated with PEASS, $[\bar{a}^{\mathrm{target}}, \bar{a}^{\mathrm{other}}]$ predicted overall quality better than the complete feature set $[\bar{a}^{\mathrm{target}}, \bar{a}^{\mathrm{artif}}, \bar{a}^{\mathrm{other}}]$. This suggests that $\bar{a}^{\mathrm{artif}}$ is not a useful feature, possibly because subjects 1) disagree strongly on which sounds qualify as "artificial noise" and/or 2) disagree strongly on how to weight this aspect into overall quality.

We obtained lower cross-validation mean squared errors with fitted mappings of the objective quality assessment methods' raw objective features compared to a linear regression of their quality predictions, except for PESQ.

V. Conclusion and Outlook

The result confirms our research hypothesis: The feature set comprising the aspect-specific quality ratings on "target preservation" and "overall disturbance of background sounds' predicts overall quality significantly better than the best set of objective features obtained from state-of-the-art objective quality assessment tools. The same applies for the set of aspect-specific quality ratings on "target preservation" and "absence of other natural sound sources".

We presume that each subjective aspect-specific feature can be predicted better than subjective overall quality by objective features, since it represents a concept on a lower abstraction level. We will investigate this by engineering objective features

[1] https://git.iem.at/stahl/subj-aspects-quality-assessment-preliminary/

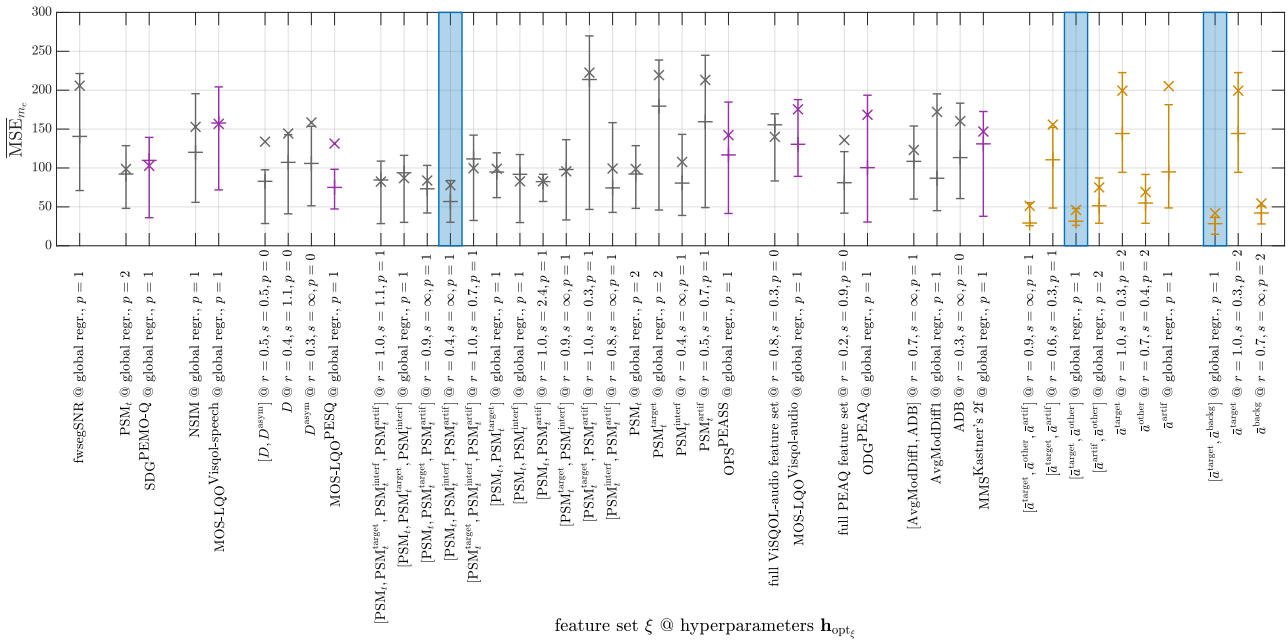

Fig. 2. Means and medians of mixture mean squared errors $\overline{\mathrm{MSE}}_{m_e}$ for different feature sets. Means are shown as crosses; medians and 95% confidence intervals are shown as bars. The feature sets with the lowest mean squared errors are highlighted.

using the collected rating data. If this presumption proves true, a new computational quality assessment method that outperforms current approaches can be designed in this way.

REFERENCES

[1] *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*. ITU-T Recommendation P.835, 2003.

[2] *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. ITU-T Recommendation P.862, 2001.

[3] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *International Conference on Independent Component Analysis and Signal Separation*, 2009, pp. 734–741.

[4] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[5] P. C. Loizou, *Speech Enhancement: Theory and Practice, Second Edition*. Boca Raton, FL: CRC Press, 2013.

[6] R. Huber and B. Kollmeier, "PEMO-Q – a new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

[7] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "ViSQOL: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–18, 2015.

[8] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "ViSQOL v3: An open source production ready objective speech and audio metric," *arXiv preprint arXiv:2004.09584*, 2020.

[9] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," in *International Conference on Latent Variable Analysis and Signal Separation*, 2012, pp. 430–437.

[10] *Method for objective measurements of perceived audio quality*. ITU-R Recommendation BS.1387-1, 2001.

[11] T. Kastner and J. Herre, "An efficient model for estimating subjective quality of separated audio source signals," in *IEEE Worksh. Applications of Signal Processing to Audio and Acoustics (WASPAA'19)*, 2019.

[12] J. Blauert and U. Jekosch, "A layer model of sound quality," in *Proc. 3rd International Workshop on Perceptual Quality of Systems (PQS 2010)*, 2010, pp. 18–23.

[13] U. Jekosch, *Voice and Speech Quality Perception: Assessment and Evaluation*. Berlin, Heidelberg: Springer, 2005.

[14] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA — a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.

[15] European Broadcasting Union. Sound quality assessment material recordings for subjective tests. Accessed: 2020-06-14. [Online]. Available: https://tech.ebu.ch/publications/sqamcd.

[16] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chime speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511.

[17] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[18] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, 1988, pp. 2578–2581 vol.5.

[19] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 196–200.

[20] E. Zwicker, H. Fastl, U. Widmann, K. Kurakata, S. Kuwano, and S. Namba, "Program for calculating loudness according to DIN 45631 (ISO 532B)," *Journal of the Acoustical Society of Japan*, vol. 12(1), pp. 39–42, 1991.

[21] *Method for the subjective assessment of intermediate quality levels of coding systems*. ITU-R Recommendation BS.1534-3, 2015.

[22] P. Kabal, "An examination and interpretation of ITU-R BS. 1387: Perceptual evaluation of audio quality," TSP Lab Technical Report, Dept. Electrical & Computer Engineering, McGill University, Tech. Rep., 2002.

[23] AudioLabs Erlangen. (2020) AudioLabs - Subjective Evaluation of Blind Audio Source Separation Database: SEBASS-DB. Accessed: 2020-08-16. [Online]. Available: https://audiolabs-erlangen.de/resources/2019-WASPAA-SEBASS.

[24] C. G. Atkeson, A. W. Moore, and S. Schaal, "Locally weighted learning," *Artificial Intelligence Review*, vol. 11, pp. 11–73, 1997.

[25] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[26] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979. [Online]. Available: http://www.jstor.org/stable/4615733