4th International Workshop on Perceptual Quality of Systems (PQS 2013)
2-4 September 2013, Vienna, Austria

Objective Quality Assessment of Target Speaker Separation Performance in Multisource Reverberant Environment

Paco Langjahr and Pejman Mowlaee

Signal Processing and Speech Communication Laboratory,
Graz University of Technology, Graz, Austria
paco.langjahr@student.tugraz.at pejman.mowlaee@tugraz.at

Abstract

Sound quality estimation of a speech enhancement or source separation system in a realistic adverse noise scenario is a challenge. In particular, the connection between results obtained by quality metrics versus those obtained from human subjective listening tests is unknown. In this paper, as the first attempt, we present results, which examine the correlation between certain objective metrics and Multi Stimulus with Hidden Reference and Anchor Point (MUSHRA) listening tests. There are a variety of applications which may benefit from the outcome of this study, since the enhancement/separation system is often used as a pre-processing stage for target applications such as - to name a few - automatic speech recognition, speech coding, speaker recognition and hearing aid devices. In all these applications the knowledge of which objective metrics reliably predicts subjective results is of great importance, as it saves the time consuming task of listening tests in future studies when developing a new enhancement system. The paper presents performance evaluation of several benchmark systems in a realistic noisy reverberant environment using the existing objective quality metrics. Subjective listening experiments which are relevant for perceptual speech quality. MUSHRA listening test is also conducted as our subjective evaluation. Comparing the objective and subjective results we present the discussion about which metrics correlate well with subjective listening tests.
Index Terms: Target speaker separation, objective metrics, multisource reverberant environment, subjective listening test.

1. Introduction

Recovering a desired target speaker signal from an adverse noisy environment has been the focus of modern science for many years now, often been addressed as a challenging and important task in speech communication systems such as mobile phones, hearing aids and automatic speech recognition systems. In this regard, there have been several challenges dealing with this topic in the past, representing different adverse noise scenarios depending on the type of background noise sources in the recording setup. In the 1st PASCAL speech separation challenge [1] the focus was on co-channel speech separation and automatic speech recognition performance without considering background noise and reverberation in the room. Although reasonable automatic speech recognition accuracies were reported, the need to extend the challenge to reverberant and more realistic scenarios led to the PASCAL CHiME speech separation and recognition challenge [2]. This time, a target speaker signal was mixed with realistic multisource reverberant environment (see Figure 2).

With the noise environment becoming more complex, the task of performance evaluation of speech enhancement methods becomes more challenging. Several objective measures and their correlation to ASR results were studied in [3, 4]. However, they did not cover human listening results in these studies and the noise in the recording setup was not from a realistic multisource reverberant environment as it comprised only three standard types. Hu and Loizou investigated the correlation between objective metrics and perceptual speech quality based on human listening tests [5, 6]. They found the Perceptual Evaluation of Speech Quality (PESQ) metric - which di Persia et al. recommended for ASR prediction - to be also the best predictor for the perceived quality. But again, the subset of the NOIZEUS corpus, which was used in both of these studies, included only four standard kinds of noise, not close to realistic scenario. Mowlaee et al. [7, 8] studied objective and subjective metrics to evaluate the performance of Blind Source Separation Evaluation (BSS EVAL) methods. The goal of the studies [7, 8] was to find out which objective metrics reliably predict the perceived speech quality and speech intelligibility obtained from listening tests. From their experiments it was observed that the Perceptual Evaluation Methods for Audio Speech Separation (PEASS) and PESQ correlate well with speech quality. Again the focus was on the task of co-channel speech separation and did not include the additional background noise or reverberation in the room. Finally, as the second PASCAL challenge, the realistic noisy database Finally in [2], the computational hearing in multisource environments (CHiME) challenge was introduced where large binaural recording data were captured in a domestic living room. The recorded data in CHiME represents speech degradations due to the reverberation in room and highly non-stationary noise sources in the background. The main focus in the CHiME challenge was to evaluate the automatic speech recognition performance of the systems participated in the challenge.

In this study we aim at estimating and linking objective and subjective performance evaluation in a realistic reverberant multisource scenario in order to discover, which of the existing quality metrics reliably predict the human listening results. To this end, we perform perceptual quality tests to finally compare between their results and the objective metric evaluation.

The rest of the paper is structured as follows: In the next

This work was partially funded by the European project DIRHA (FP7-ICT-2011-7-288121) and by the K-Project AAP in the COMET (Competence Centers for Excellent Technologies) programme with joint support from BMVIT, BMWFJ, Styrian Business Promotion Agency (SFG), and the Government of Styria ("Abt. 3: Wissenschaft und Forschung" as well as "Abt. 14: Wirtschaft und Innovation"). The programme COMET is managed by the Austrian Research Promotion Agency (FFG).

10.21437/PQS.2013-16

section we present some background information about the metrics and benchmark systems that have been evaluated as well as the dataset and the experimental setup. After that, the subjective and objective speech quality results will be presented, followed by a conclusion about which quality metric(s) reliably predicts the human listening results.

2. Background

2.1. Performance Evaluation: a multi-dimensional optimization problem

In general, sound quality in speech communication systems can be judged according to several aspects. In fact, the task of performance evaluation of a signal enhancement method is a multidisciplinary one. Figure 1 shows four different ways to assess the performance of a signal enhancement method: 1) using instrumental metrics to report the perceived speech quality e.g. PESQ [9] or SNR-based measures [10], 2) using speech intelligibility metrics e.g. STOI [11] or SNRloss [12]. 3) Conducting subjective listening tests e.g. measured by MUSHRA test [13] or speech intelligibility test described in [14]. 4) Measuring the word recognition accuracy as the outcome of an automatic speech recognizer applied to the enhanced speech signal. For the first two methods we require the access to the clean speech signal. The third method (subjective listening tests) is the most reliable one to evaluate the performance of the system, as the human listener is in fact the end user of a speech communication system.

It is clear that the criterion used in instrumental metric, automatic speech recognition are arguably different than the one used for automatic speech recognition as a classifier. The methods in the third and the fourth group both calculate the number of correctly recognized words by a human listener or a machine, so they are somehow connected, but to which degree is unknown. Furthermore, perceived quality does not always correlate with Speech Intelligibility. Therefore, a respectable number of different measures have been found in preceding studies. The paper of Möller et al. provides a comprehensive overview about the state-of-the-art objective metrics that have been and will be developed in the past and the future [15].

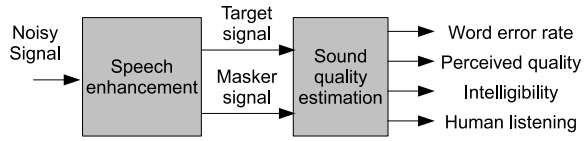

Figure 1: Block diagram showing four different ways for evaluating the performance of a speech enhancement or source separation method in a multisource reverberant environment.

2.2. Objective Metrics

The following objective metrics were employed to evaluate the speech enhancement and separation performance of each benchmark method: Perceptual Evaluation of Speech Quality (PESQ), Blind Source Separation Evaluation (BSS-EVAL) including Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR) and Signal to Artifacts Ratio (SAR), Perceptual Evaluation Methods for Audio Source Separation (PEASS) including Overall Perceptual Score (OPS), Target-related Perceptual Score (TPS), Interference-related Perceptual Score (IPS) and Artifacts-related Perceptual Score (APS), Cepstral Distance

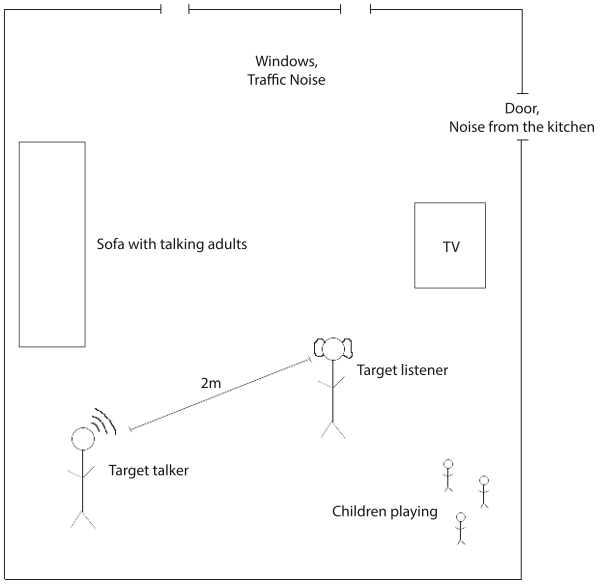

Figure 2: Plan of the CHiME recording setting showing location of the target speaker and the most significant noise sources.

(CD), frequency-weighted segmental SNR, Segmental SNR (SSNR) and overall SNR.

2.2.1. Metrics for Speech Enhancement

As our first metric we employed in addition to conventional SNR the Segmental SNR (SSNR) [16] which is defined as:

$$d_{\text{SSNR}}(\boldsymbol{s_m}, \hat{\boldsymbol{s}}_{\boldsymbol{m}}) = 10log_{10}\frac{||\boldsymbol{s_m}||^2}{||\hat{\boldsymbol{s}}_{\boldsymbol{m}} - \boldsymbol{s_m}||^2}, \quad (1)$$

where $||.||$ is the 2-norm defined as $||\text{x}|| = \sqrt{(\sum_{n=1}^{N} x[n]^2)}$.

Following the notation in [3], the original signal is denoted as $\boldsymbol{s_m}$ and the enhanced signal is $\hat{\boldsymbol{s}}_{\boldsymbol{m}}$, both of M samples. Frame m of length N of the original signal is defined as $\boldsymbol{s_m} = [s[mQ], ..., s[mQ + N - 1]]$,where Q is the step size of the window in a short-time analysis, and with analogous definition for the corresponding frame of the separated signal.

Also evaluated is the performance of the benchmark methods in terms of Cepstral Distance (CD). As reported in [17], CD is defined as:

$$d_{\text{CD}}(\boldsymbol{s_m}, \hat{\boldsymbol{s}}_{\boldsymbol{m}}) = \frac{10}{log10}\sqrt{2\sum_{l=1}^{p}(c_m[l] - \hat{c}_m[l])^2}, \quad (2)$$

where c_m and $\hat{\text{c}}_m$ are the coefficients of the vectors corresponding to a frame of the original signal and the enhancement result, and p is the order of the linear prediction model. CD is known as an effective evaluation measure for coding distortion and other nonlinear distortions such as quadratic and logarithmic distortion.

Another quite common objective metric is the frequency-weighted segmental SNR (fwSNRseg), which was computed

using the following equation [18]:

$$\text{fwSNR}_{\text{seg}} = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K} W(j,m)log_{10}(\frac{S(j,m)^2}{(S(j,m)-|\hat{S}(j,m)|)^2})}{\sum_{j=1}^{K} W(j,m)}, \quad (3)$$

where $W(j,m)$ is the weight applied to the jth frequency band, K is the number of bands, M is the total number of frames in the signal, $S(j,m)$ is the clean signal spectrum in the jth frequency band at the mth frame, weighted by a Gaussian-shaped window. $|\hat{S}(j,m)|$ and in the weighted enhanced signal spectrum in the same band.

Perceptual evaluation of speech quality (PESQ), is a common sophisticated metric proposed by the ITU-T recommendation for the purpose of speech quality assessment of coded speech with telephony quality. However, it was found as a good metric to predict the speech enhancement [6] or separation [7] performance.

The definition as can be found in [9] is as follows:

$$\text{PESQ} = a_0 + a_1 D_{ind} + a_2 A_{ind}. \quad (4)$$

where D_{ind} is the average asymmetrical disturbance and A_{ind} is the asymmetrical disturbance. The parameters a_0,a_1 and a_2 were not modified and the values are according to [9] a_0=4.5, a_1=-0.1 and a_2=-0.0309.

2.2.2. Blind Source Separation Metrics

To have comparisons with previous studies, here we include two BSS metrics: BSS EVAL toolkit [19] and PEASS toolkit [20]. BSS EVAL covers the objective metrics SDR, SIR and SAR.

- Signal-to-Distortion Ratio (SDR): measures the amount of distortion introduced in the output signal and is defined as the ratio between the energy of the clean signal and that of the distortion.
- Signal-to-Interference Ratio (SIR): is defined as the ratio between the power of the target signal and that of the interference signal and is used to measure the amount of undesired interference still remaining in the separated signal.
- Signal-to-Artifacts Ratio (SAR): measures the quality in terms of absence of artificial noise.

The PEASS toolkit includes OPS, IPS, APS and TPS.

- Overall Perceptual Score (OPS): measures how close the separated signal is to the clean reference signal.
- Target-related Perceptual Score (TPS): is defined as how well the target is preserved in the separated signal compared to the clean signal.
- Interference-related Perceptual Score (IPS): measures the interference cancellation in the separated signal.
- Artifact-related Perceptual Score (APS): shows the quality of the separated signal in terms of having no artifacts.

The BSS Evaluation Metrics are defined in [10] as:

$$\text{SDR} = 10log_{10}\frac{||\boldsymbol{s_{target}}||^2}{||\boldsymbol{e_{interf}} + \boldsymbol{e_{noise}} + \boldsymbol{e_{artif}}||^2}. \quad (5)$$

$$\text{SIR} = 10log_{10}\frac{||\boldsymbol{s_{target}}||^2}{||\boldsymbol{e_{interf}}||^2}. \quad (6)$$

$$\text{SAR} = 10log_{10}\frac{||\boldsymbol{s_{target}} + \boldsymbol{e_{interf}}||^2}{||\boldsymbol{e_{noise}}||^2}. \quad (7)$$

In Equations (5) to (7), $\boldsymbol{s_{target}}$ denotes a version of the true desired source modified by an allowed distortion and $\boldsymbol{e_{interf}}, \boldsymbol{e_{noise}}$ and $\boldsymbol{e_{artif}}$ are the interferences, noise and artifacts error terms.

The PEASS metrics are given in [20]. They are derived from the distortion components decomposed in the BSS Evaluation measures. The salience of the overall distortion and each distortion component is assessed using auditory model-based metrics to overcome the frequency dependency of i.e. artifacts or auditory masking. Following other measures, a nonlinear mapping function is applied to combine these salience features into a single scalar measure for each grading task. These scalar values are the Overall Perceptual Score (OPS), the target-related perceptual score (TPS), the Interference-related Perceptual Score (IPS) and the Artifacts-related Perceptual Score (APS).

3. Results

3.1. Database, Benchmark Methods and System Setup

3.1.1. Speech Database

In our experiments, we extract the noisy signals from the CHiME[1] corpus [21]. It consists of 57 recording sessions made with the binaural B&K HATS in two domestic rooms, a lounge and a kitchen with a reverberation time of 300 milliseconds. Figure 2 shows the recording setting of CHiME challenge. The recording sessions give a sound material of around 14 hours with many different noise sources like two adults, two children, TV, footsteps, toys, games console and some traffic noise from outside. Similar to the 1st PASCAL speech separation challenge [1], CHiME corpus consists of 600 utterances recorded from 34 speakers (18 male and 16 female). The utterances are designed by the combination of a grid of possible words as shown in Table 1. Each sentence includes a command, a color, a preposition, a letter, a digit and an adverb. The listener has to recognize the color, the letter and the digit, whereas the other words are "fillers" to create some variation in contexts for the neighboring key words. The mixtures of the spoken utterances with noisy background are made such that the placement in time of the Grid sentences in the noise background is controlled to produce 600 sentences with 6 different Signal to Noise Ratios (SNR), ranging in -6dB, -3dB, 0dB, 3dB, 6dB and 9dB. The nature of the background noise used in CHiME recording varies from backgrounds highly non-stationary energetic events at low SNRs (-6 dB) to fairly stationary ambient noise at high SNRs (9 dB).

Since our study includes listening tests and listener fatigue is always a difficulty with a lot of data, we had to limit the CHiME data for the subjective tests. The SiSEC Subset known from the CHiME challenges provides already a limitation with

Table 1: Sentence structure of the Grid corpus.

command	color	preposition	letter	digit	adverb
bin	blue	at	A-Z	1-9,zero	again
lay	green	by	excluding W		now
place	red	in			please
set	white	with			soon

[1] http://spandh.dcs.shef.ac.uk/chime_challenge/

4 clips per SNR and further reduction of the data was archived by choosing one female and one male speaker per SNR out of the SiSEC Subset (in total 12 clips per method) for the listening tests. For the sake of consistency, we exploited the same dataset for the objective metric evaluation.

3.1.2. Benchmark Methods

To explore the relationship between the objective and subjective results we chose three top performing systems participated in the CHiME 1 challenge. The three systems evaluated in this study are Weninger et al. [22], Koldovský et al. [23] and Hurmalainen et al. [24]. These systems employ different signal processing strategies, which are target enhancement, robust feature extraction, robust decoding and trained noise model [2].

In the first strategy *target enhancement* the noisy input signal is represented in the time-frequency domain and a linear filter is applied to each time-frequency bin. This linear filter is mostly a combination of a spatial filter resulting from a fixed or an adaptive beamformer with a spectral filter such as high-pass, low-pass, Wiener or binary or softmask.

The second strategy called *robust feature extraction* is aimed to extract features robust to background noise by either robust features such as Gammatone Frequency Cepstral Coefficients (GFCC) or feature transformations such as Maximum Likelihood Linear Transformation (MLLT).

In the third strategy *robust decoding* the features are transformed into a sequence of words by using a conventional HMM-GMM recognizer, which is improved with either multi-condition training, robust training, noise-aware decoding or system combination. The *model combination* jointly decodes the target and the background without any target enhancement front-end. The System of Hurmalainen et al. [24] uses robust features, robust decoding and trained noise model, while the system of Koldovský et al. [23] only uses target enhancement. The System of Weninger et al. [22] uses all of the four strategies.

3.1.3. Subjective Test Setup

The listening test was held in a quiet room with closed AKG K270 Studio Headphones and the manual to the test were presented to the listeners via an interactive MATLAB GUI where the participants could read the instructions whenever they wanted. Following the MUSHRA Standard proposed in ITU-R BS.1534-1 [13], we included a full bandwidth Hidden Reference and an additional Anchor Point, which was a low pass @3.5 kHz filtered version of the reference signal. The number of listeners was set to eight, as this was expected to result in short enough confidence intervals[2].

3.2. Objective and Subjective Results

The mean opinion score (MOS) result obtained from MUSHRA test is shown in Figure 4. Results are reported for the unprocessed noisy signal as well as the three benchmark systems that participated in the CHiME challenge described in section 3.1.2. The figure shows the results averaged over all listeners and all clips. The error bars show the 95% confidence intervals. Observing the barplot shows up the following ranking: The method of Hurmalainen et al. [24] is the best method in terms of perceptual quality, then follows the method of Koldovský et al. [23],

[2]Matlab implementations for subjective listening test are available at `http://www.audis-itn.eu/wiki/Matlabcodes`

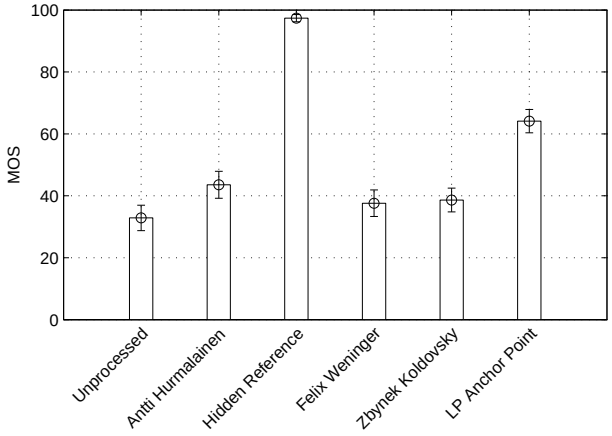

Figure 4: MUSHRA results for different separation methods averaged over all excerpts and listeners. Error bars indicate 95% confidence interval

on the third place is the method of Weninger et al. [22] and the unprocessed signal has the lowest ranking, as expected. T-tests were needed to prove the significance of this ranking. With a p-value of 0.05 the null hypothesis could not be rejected for the test between Weninger et al. [22] and Koldovský et al. [23], but with p=0.06 the significance between all three methods and the unprocessed signal could be proved. This raises an assumption that with a few more listeners the confidence intervals would have shortened enough to show the clear ranking, which can already been estimated in the barplot. A closer look at the results in Table 2 reveals, that the metrics which promise good prediction quality are PESQ and OPS, because in these metrics the ranking is the same as in the subjective listening test.

Table 2: Objective Results on SiSEC Subset of CHiME, averaged over all clips and SNR levels. The best result in each metric is highlighted in bold face. UP: Unprocessed, AH: Antti Hurmalainen, FW: Felix Weninger, ZK: Zbyněk Koldovský

	UP	AH	FW	ZK
SDR	-5,05	**0,04**	-0,47	-0,64
SIR	-2,41	9,07	**10,33**	6,19
SAR	**6,17**	1,29	0,40	1,86
OPS	25,49	**28,87**	25,57	26,67
TPS	**35,32**	34,20	31,41	34,46
IPS	38,96	65,95	**76,23**	71,61
APS	**41,63**	30,37	24,24	30,59
Cep. Dist.	5,57	**4,34**	5,05	5,12
SNR	-5,77	-2,65	-3,77	**-2,10**
Seg. SNR	-5,94	-2,74	-3,98	**-2,66**
Freq. SNR	3,23	**3,76**	3,21	3,24
PESQ	1,65	**2,09**	1,84	1,88

3.3. Paired Test Results

To find out a reliable predictor among the objective metrics for subjective results presented in Fig. 4, we conduct the paired t-tests. As a prerequisite for reliable prediction, the result of a metric in Table 3 has to correlate with the subjective result, which means in this case that the null hypothesis of every t-test

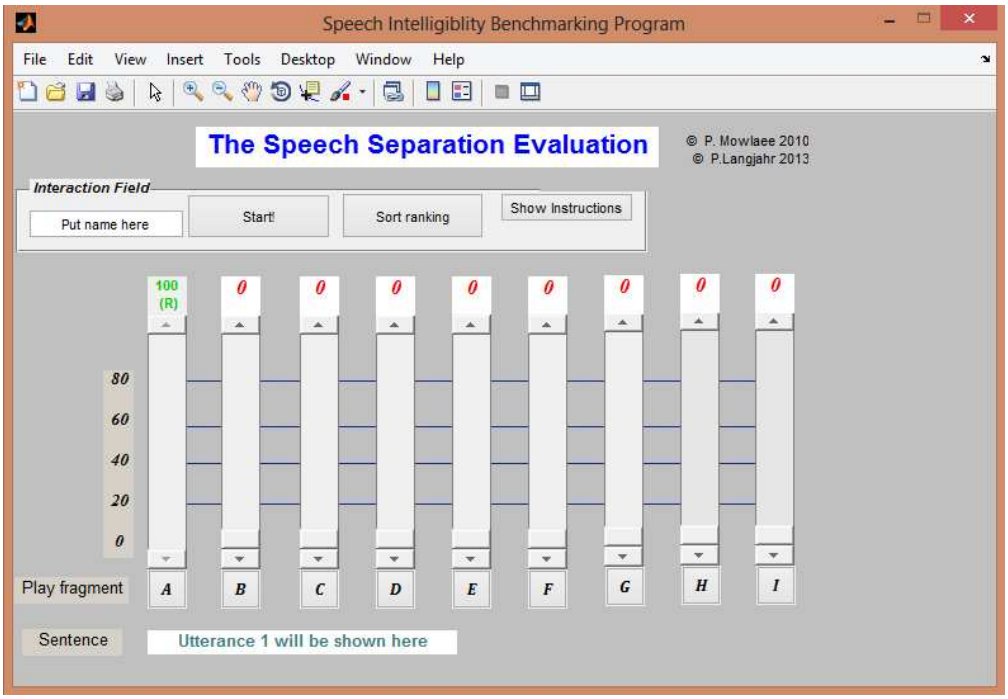

Figure 3: MATLAB GUI for the MUSHRA test.

has to be rejected. This is not fulfilled by any of these two metrics. Even if the methods of Weninger et al. [22] and Koldovský et al. [23] are assumed to be not significantly different in their results (as is shown by the listening test under the assumption of a sufficient number of listeners), none of the metrics follow the subjective ones in Table 3, which would mean a one in all columns except for the last one. PESQ, previously known as a good predictor of the perceptual speech quality [6], also fails in this scenario for multisource reverberant environment, as it shows better results for Koldovský et al. than for Weninger et al. [22] and the t-test rejects the null hypothesis for a p-value lower than 0.03. Therefore, a new or revised metric concerning perceptual speech quality needs to be found.

4. Conclusion

In this study we aimed at investigating, whether for multisource reverberant environment there are objective metrics, which reliably predict the results of subjective listening tests. This would be a highly requested advantage for future studies, since it could save a lot of time. We presented the objective evaluation results for 14 metrics on the dataset of a part of the SiSEC Dataset. A subjective listening for speech quality using MUSHRA was performed. The results were presented and compared with the objective results of the SiSEC Dataset. No metric was able to predict the human perceptual quality results reliably. Although PESQ and OPS matched the ranking of the mean values of the subjective results, the statistical significance test could not prove the reliability of this result.

Future studies should focus on the revision of these metrics or proposing new metrics to address the issue of speech quality estimation in a realistic multisource reverberant environment, otherwise extensive listening tests will still be unavoidable for meaningful results in speech quality estimation of a new proposed algorithm.

The ultimate goal in a single-channel speech enhancement is to recover the time-domain clean signal, i.e., both amplitude and phase spectra are required to be estimated. Existing quality metrics are defined based on a l_2-norm as squared error metric, to evaluate the performance of an enhancement system. Therefore, a novel quality metric which reflect the amplitude and phase estimation accuracy is required. In particular, to evaluate the performance of recent phase-aware signal enhancement methods one require some phase-based metrics, to measure the performance of phase enhanced methods recently proposed in [25–29]. Our ongoing study is directed toward discovering new phase based quality metrics which predict the subjective listening results obtained by the proposed phase-aware speech enhancement method e.g. presented in [27].

5. Acknowledgement

We would like to thank Dr. Felix Weninger, Dr. Antti Hurmalainen and Dr. Zbyněk Koldovský for their assistance in sharing their data, to be included as our benchmark methods.

6. References

[1] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, 2010.

[2] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech and Language, Special Issue on Multisource Environments*, vol. 27, no. 3, pp. 621–633, Jan. 2012.

[3] L. D. Persia, M. Yanagida, H. L. Rufiner, and D. Milone, "Objective quality evaluation in blind source separation for speech recognition in a real room," *Signal Processing*, vol. 87, no. 8, pp. 1951–1965, Aug. 2007.

Table 3: t-test results over the objective evaluation of the SiSEC subset of CHiME. UP: Unprocessed, AH: Antti Hurmalainen, FW: Felix Weninger, ZK: Zbyněk Koldovský

	UP vs. AH	UP vs. FW	UP vs. ZK	AH vs. FW	AH vs. ZK	FW vs. ZK
SDR	1	1	1	0	0	0
SIR	1	1	1	0	1	1
SAR	1	1	1	1	0	1
OPS	0	0	0	1	1	0
TPS	0	0	0	0	0	0
IPS	1	1	1	1	1	1
APS	1	1	1	1	0	1
Cep. Dist.	1	0	1	1	1	0
SNR	1	0	1	1	1	1
Seg. SNR	1	1	1	1	0	1
Freq. SNR	1	0	0	1	1	0
PESQ	1	0	1	1	1	0

[4] L. Di Persia, D. Milone, H. L. Rufiner, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Process.*, vol. 88, no. 10, pp. 2578–2583, Oct. 2008.

[5] Y. Hu and P.C. Loizou, "Subjective comparison of speech enhancement algorithms," in *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2006, pp. 153–156.

[6] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[7] P. Mowlaee, R. Saeidi, M.G. Christensen, and R. Martin, "Subjective and objective quality assessment of single-channel speech separation algorithms," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, march 2012, pp. 69 –72.

[8] P. Mowlaee, R. Saeidi, M.G. Christensen, Zheng-Hua Tan, T. Kinnunen, P. Franti, and S.H. Jensen, "A joint approach for single-channel speaker identification and speech separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2586 –2601, nov. 2012.

[9] 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," *Elsevier speech communication*, vol. 2, pp. 749–752, Aug. 2001.

[10] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462 –1469, Jul. 2006.

[11] C.H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, September 2011.

[12] J. Ma and P.C. Loizou, "SNR loss: A new objective measure for predicting the intelligibility of noise-suppressed speech," *Speech Communication*, vol. 53, no. 3, pp. 340–354, 2011.

[13] International Telecommunication Union, "Method for the subjective assessment of intermediate quality level of coding systems," *ITU-R BS.1534-1*, 2001-2003.

[14] Jon Barker and Martin Cooke, "Modelling speaker intelligibility in noise," *Speech Commun.*, vol. 49, no. 5, pp. 402–417, May 2007.

[15] S. Möller, W.-Y. Chan, N. Ct, T. H. Falk, A. Raake, and M. Wltermann, "Speech quality estimation: Models and trends," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 18–28, 2011.

[16] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *In proceedings of the international conference on speech and language processing*, 1998.

[17] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," *Journal on Selected Areas in Communications*, vol. 6, no. 2, Feb. 1988.

[18] J. Ma, Y. Hu, and P.C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *Journal of the Acoustic Society of America*, vol. 125, no. 5, May 2009.

[19] C. Févotte, R. Gribonval, and E. Vincent, "BSS_EVAL Toolbox User Guide – Revision 2.0," Tech. Rep., 2005.

[20] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, , no. 99, pp. 1, 2011.

[21] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," in *Proc. Interspeech*, 2010.

[22] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The munich 2011 chime challenge contribution: NMF-BLSTM speech enhancement and recognition for reverberated multisource environments," in *Proc. Machine Listening in Multisource Environments (CHiME 2011), Florence, Italy*, 2011.

[23] Z. Koldovsky, J. Malek, J. Nouza, and J. Balik, "Chime data separation based on target signal cancellation and noise masking," in *Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, 2011.

[24] A. Hurmalainen, J.F. Gemmeke, and T. Virtanen, "Modelling non-stationary noise with spectral factorisation in automatic speech recognition," *Computer Speech and Language*, vol. 27, no. 3, pp. 763–779, 2013.

[25] P. Mowlaee, M. Watanabe, and R. Saeidi, "Show & tell: Phase-aware single-channel speech enhancement," in *14th Annual Conference of the International Speech Communication Association*, 2013.

[26] P. Mowlaee, R. Saeidi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," in *Proceedings of the International Conference on Spoken Language Processing*, 2012.

[27] P. Mowlaee and R. Saeidi, "On phase importance in parameter estimation in single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2013, pp. 7462 – 7466.

[28] P. Mowlaee and R. Martin, "On phase importance in parameter estimation for single-channel source separation," in *The International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012.

[29] M. Watanabe and P. Mowlaee, "Iterative sinusoidal-based partial phase reconstruction in single-channel source separation," in *Proceedings of the International Conference on Spoken Language Processing*, 2013.