PERFORMANCE COMPARISON OF INTRUSIVE AND NON-INTRUSIVE INSTRUMENTAL QUALITY MEASURES FOR ENHANCED SPEECH

Anderson Avila[1], *Benjamin Cauchi*[2,4], *Stefan Goetze*[2,4], *Simon Doclo*[3,4], *Tiago Falk*[1]

[1]INRS-EMT, University of Quebec, Montreal, QC, Canada
[2]Fraunhofer IDMT, Project Group Hearing, Speech and Audio Technology, Oldenburg, Germany
[3]University of Oldenburg, Dept. of Medical Physics and Acoustics, Oldenburg, Germany
[4]Cluster of Excellence Hearing4all, Oldenburg, Germany

ABSTRACT

Instrumental quality prediction of speech processed by enhancement algorithms has become crucial with the proliferation of far-field speech applications. To date, while several instrumental measures have been proposed and standardized, their performance under a wide range of acoustic conditions and enhancement algorithms is still unknown. This paper aims to fill this gap. Specifically, the performance of eleven instrumental measures are compared; four are non-intrusive measures, i.e. not requiring a clean reference signal, and seven intrusive. Simulated and recorded speech under four different acoustic conditions involving varying levels of reverberation and noise are explored, as well as processed by three single- and multi-channel enhancement algorithms. Experimental results show that a recently developed non-intrusive measure called $\text{SRMR}_{\text{norm}}$ outperforms all other considered measures in terms of overall quality prediction. The well-known PESQ measure, in turn, showed to better predict the perceived amount of reverberation, followed by $\text{SRMR}_{\text{norm}}$. These results are promising, as the latter measure does not require access to a clean reference signal, thus has the potential to be used for enhancement algorithm optimization in real-time.

Index Terms— Speech quality, perceptual evaluation, instrumental measures, microphone array, speech enhancement

1. INTRODUCTION

In hands-free speech applications, such as voice-controlled systems or hearing devices, speech is captured via a far-field microphone or microphone array. Under such condition, the speech signal is typically degraded by both ambient noise and room reverberation. Depending on the room properties (e.g., large reverberation time) and environmental conditions (e.g., noisy office setting), this can severely degrade the performance of automated speech technologies, as well as the perceived quality and intelligibility for human listeners. To overcome this limitation, single- and multi-microphone based speech enhancement algorithms have been proposed.

The most common single-microphone approach relies on a technique termed spectral enhancement (SE), which is based on estimators of the spectral amplitude of the clean signal [1] using an estimate of the power spectral density (PSD) of the interference to be suppressed [2], be it noise (e.g., [3, 4]) and/or reverberation (e.g., [5, 6]). In turn, when multiple microphones are available, spatial information can be exploited to improve speech enhancement performance [7]. The classical microphone array based enhancement algorithm is beamforming, which uses an estimate of the direction of arrival (DOA) of the target speech source to suppress interferences from different directions [8]. Alternative strategies in addition take into account the spectro-temporal information of the speech and noise sources. Representative examples include the multichannel Wiener filter [9], which corresponds to the combination of a minimum variance distortionless response (MVDR) beamformer with a Wiener gain, as well as other combinations of beamforming with spectral enhancement (e.g., [10, 11]).

While existing enhancement algorithms are capable of significantly reducing noise and reverberation from the captured speech signal(s), suboptimal algorithm parameter tuning may lead to the introduction of unwanted artefacts, which in turn, can degrade speech intelligibility, quality, and acceptability [12]. As examples, spectral enhancement algorithms are known to introduce so-called musical noise artefacts [13] and badly tuned beamformers can mistakenly suppress the target signal, as opposed to the interferer. To date, there is no reliable means of automatically (i.e., instrumentally) measuring the quality or intelligibility of an enhanced speech signal, such that enhancement algorithm parameters could be tuned in real-time to maximize the performance.

Instrumental measurement of the quality of enhanced speech signals is an area of growing interest, particularly given the recent advances with far-field speech technologies, such as automatic speech recognition. Instrumental measures can be classified as either intrusive or non-intrusive, depending on the need for a (clean) reference signal [14]. While existing instrumental measures have been validated in the past, their focus has been rather limited. For example, in [15], only reverberant speech and speech processed by a simple delay-and-sum beamformer was explored. In [11], in turn, only reverberant speech and signals corrupted by reverberation and noise were tested. In [16], speech processed by several noise reduction algorithms was used and, more recently in [17], single-channel dereverberation algorithms were tested, but assumed perfect knowledge of room parameters, such as the true room impulse response (RIR). As such, a more complete performance comparison is still needed where different enhancement algorithms are used under varying acoustic conditions. Moreover, evaluating the performance of existing instrumental measures as correlates of rating scales different from the well-known mean opinion score (MOS) is also lacking. As suggested by the International Telecommunication Union (ITU-T), this is critical for enhancement algorithms [18].

In this paper, the performance of eleven instrumental measures is evaluated as correlates of the perceived speech quality **and** perceived level of reverberation in signals processed by three speech enhancement strategies, namely MVDR beamforming, single-channel

This research has been partially funded by the Deutscher Akademisher Auschdienst (DAAD) and by a Google Faculty Research Award.

978-1-5090-2007-2/16/$31.00 ©2016 Crown

spectral enhancement, and a combination of the two approaches, as proposed in [11], under four different acoustic conditions.

2. INSTRUMENTAL MEASURES

In this section, a brief description of the considered non-intrusive and intrusive instrumental measures is given.

2.1. Non-intrusive measures

2.1.1. ANIQUE+

ANIQUE+ is a measure standardized by the American National Standards Institute for telephone-band speech. The measure expresses speech distortion as the sum of three specific distortion types, namely individual frame distortion (based on articulation analysis), mute distortion (detects abrupt starts and stops), and non-speech distortion, which are then mapped to a final quality rating using an artificial neural network (ANN). The ANN was trained on large amounts of multi-lingual telephone narrowband speech data under a wide range of telecommunication network conditions, including standard and nonstandard speech codecs, transcoding, channel errors, packet loss and its concealment, environmental noise at the sending side, time-varying delay, and acoustic coupling at the sending side. The reader is referred to [19] for complete details.

2.1.2. ITU-T Rec. P.563

The ITU-T Recommendation P.563 is also a non-intrusive standard measure for telephone-band speech. As ANIQUE+, it combines three basic principles for evaluating distortions. First it models the human voice production system as a series of vocal tract tubes and abnormal values are considered to be distortions. Second, it constructs a pseudo-clean signal from its noisy counterpart in order to apply an intrusive measure in order to obtain an intermediate quality index. Third, specific distortions encountered in voice channels, such as temporal clipping, robotization, and noise are detected and quantified. The different distortion types are then ranked and a distortion-dependent weighted linear mapping is applied to estimate the final MOS rating. Complete details can be encountered in [20].

2.1.3. Speech-to-Reverberation Modulation Ratio, SRMR

The so-called SRMR measure relies on the principle that the modulation energy of clean speech is generally concentrated in lower modulation frequencies (below 20 Hz) while room acoustic artefacts typically arise in higher modulation frequencies beyond 20 Hz. As such, the metric computes the ratio of low to high modulation energy after an auditory model is applied based on a 23-channel gammatone cochlear filterbank and an 8-channel modulation filterbank to emulate the human hearing system. The model has been shown to accurately characterize room acoustics, as well as the quality and reverberation level of reverberant speech and speech processed by a simple delay-and-sum beamformer [15, 21].

2.1.4. SRMR$_{norm}$

Recently, an extended version of the SRMR measure was proposed in order to reduce the variability caused by the effects of pitch and speech content [21]. In order to reduce pitch effects, the frequency range of the modulation filters was reduced from 4-128 Hz in the original SRMR implementation to 4-40 Hz. Second, in order to reduce the sensitivity to spoken content, a per-frame energy thresholding scheme was implemented where only frames below 30 dB of the maximum were used. SRMR$_{norm}$ was shown to reduce intra- and inter-speaker variability and to better estimate the intelligibility level of speech under reverberation, noise, and reverberation-plus-noise conditions [21]. In this paper, a first exploration into the use of this measure for enhanced speech is taken.

2.2. Intrusive instrumental measures

2.2.1. Perceptual Evaluation of Speech Quality, PESQ

ITU-T Recommendation P.862, also known as Perceptual Evaluation of Speech Quality, is the most widely-used intrusive instrumental measure available for telephone-band speech [22]. As an intrusive measure, PESQ first relies on a time alignment algorithm in order to directly compare the reference and processed signals. The signals are then transformed to an internal representation that is analogous to the psychophysical representation of audio signals in the human auditory system by means of perceptual frequencies and compressive loudness scaling. The difference in internal representations of the degraded and reference speech signals is then calculated, representing the audible difference between the two signals. Lastly, a cognitive model evaluates audible errors by computing two types of noise disturbances for individual time-frequency bins, namely asymmetrical and symmetrical disturbances. The predicted mean opinion score (MOS) is calculated as a weighted linear combination of these disturbances. More details can be found in [22, 23].

2.2.2. Perceptual Objective Listening Quality Assessment, POLQA

Recently, PESQ was superseded by the ITU-T Recommendation P.863 [24], also known as Perceptual Objective Listening Quality Assessment (POLQA). The core of POLQA is very similar to PESQ [25], but with some critical changes needed for evolving speech technologies. For example, improved time alignment was implemented in order to compensate for distortions seen in recent packet loss concealment strategies. In addition to the symmetric and asymmetric disturbances used within PESQ, POLQA includes an additional noise analysis module and a reverberation analysis module. These parameters are then combined within a cognitive mapping module to estimate MOS. Unlike PESQ, POLQA has also been designed to cover a wide range of speech technologies, ranging from narrowband to super-wideband (50-14,000 Hz) [24]. Given these additional modules, it is expected that POLQA performs better than PESQ for enhanced speech.

2.2.3. Normalized Covariance Metric, NCM

The NCM is based on the covariance between auditory-inspired envelopes of the clean and processed speech signals [26]. This is achieved by means of a gammatone filterbank to emulate cochlear processing, followed by a Hilbert transform. The NCM is shown to be a good estimator of quality and intelligibility of reverberant speech for impaired listeners with and without hearing devices [27].

2.2.4. Short-Time Objective Intelligibility, STOI

As with the NCM, the STOI measure is based on the covariance between the temporal envelopes of the clean and processed speech and assumes that both signals are time-aligned. The main difference is the fact that the STOI metric computes this covariance over short time frames and then aggregates the per-frame disturbances into a final quality rating. Unlike NCM, the speech signals are decomposed by a one-third octave filterbank, followed by level normalization and clipping. More details can be found in [28].

2.2.5. *Coherent Speech to Intelligibility Index, CSII*

The CSII was originally proposed for hearing aid devices [29]. Assuming the clean and processed signals are time-aligned, they are divided into three different amplitude regions: high-level segments falling above the overall root-mean-square level, mid-level segments between 0 and 10 dB, and low-level segments between 10 and 30 dB. It was shown that the concentration of vowels (and peak clipping) will be found in the high-level segments, vowel-consonant transitions in the mid-level segments, and the low-level segments will contain consonants, pauses, additive noise and center clipping [26]. These disturbances are mapped to a final intelligibility rating.

2.2.6. *Log Likelihood Ratio, LLR*

The log likelihood ratio is a traditional intrusive measure that relies on the linear prediction model of human speech production. This measure is based on the distance between the linear prediction coefficients (LPC) attained from the clean and processed speech signals. Details on this measures can be found in [30].

2.2.7. *Itakura-Saito distance, IS*

As an LPC-based measure, the Itakura-Saito (IS) distance measures the difference between the spectral envelope of the clean and processed signal [31, 30]. The distance is computed frame by frame, with low values indicating similarity between original and processed signal. A detailed explanation of the IS measure is provided in [31].

3. EXPERIMENTAL SETUP

In this section, we describe the dataset used to compare the different instrumental measures, as well as the figures-of-merit used.

3.1. Database description

The data used in our experiments was the evaluation set of the 2014 IEEE REVERB challenge [12], which consists of a large corpus of speech corrupted by varying levels of reverberation and noise. All recordings were made with a sampling frequency of 16 kHz with a circular microphone array with 20 cm diameter and 8 equidistant microphones. The corpus is divided into two sets: one simulated and another comprised of real recorded data. The simulated set is composed of clean speech signals taken from the WSJ-CAM0 corpus [32], which were convolved with RIRs recorded in three different rooms and to which measured noise at a fixed signal-to-ratio (SNR) of 20 dB was added. The real recorded set, in turn, is composed of utterances from the MC-WSJ-AV corpus [33] and contains speech recorded in a room in the presence of noise.

In our experiments, the medium room in the simulated set and the large room in the recorded set were used. For each room, two distances (denoted by "near" and "far") between the target speaker and the center of the microphone array were used. Table 1 describes the reverberation time and distances of these four "conditions".

Table 1: Description of the four acoustic conditions tested. Column labelled T_{60} corresponds to the reverberation time of the room.

Set	Room	T_{60} [ms]	Distance [cm]	Label
Simulated	medium	500	50	S2, near
			200	S2, far
Real	large	700	100	R1, near
			250	R1, far

Three speech enhancement algorithms were applied to the degraded speech signals, namely the MVDR beamformer (termed MVDR henceforth), single-channel spectral suppression (SS) applied to the first channel of the array (termed SE_3), and a combination of the two algorithms where SS has been applied to the output of the MVDR beamformer, as proposed in [11] (termed MVDR+SE_3). The MVDR beamforming coefficients are computed using an online-estimated noise coherence matrix and the DOA of the target speech source was estimated using the multiple signal classification (MUSIC) algorithm [34]. The SE scheme, in turn, relies on estimates of the noise and reverberant power spectral densities computed using a modified version of the minimum statistics (MS) estimator proposed in [35] and the statistical model of the RIR proposed in [5]. Since all parameters required for of single- and multi-channel algorithms are estimated online, errors are expected, thus unwanted artefacts are likely to be introduced; this was later verified via listening tests (more details below). As such, it is expected that the results presented in this paper can be generalized to real applications with alternate enhancement schemes.

In order to test the accuracy of the instrumental measures, ground truth subjective ratings are needed. Here, a listening test was conducted using the multiple stimuli test, as described in [36]. Twenty one self-reported normal-hearing listeners participated in the listening test. Participants were presented with the degraded speech signals, a hidden reference, an anchor, and the enhanced signals using the three different enhancement algorithms described above. The hidden reference was the anechoic speech signal in the case of conditions "S2, near" and "S2, far" (see Table 1) and the signal recorded by a headset microphone in the case of conditions "R1, near" and "R1, far". The anchor, in turn, consisted of the first microphone signal of the array, low-pass filtered with a cut-off frequency of 3.5 kHz. The listening test was conducted in a soundproof booth and participants listened to diotic signals through headphones (Seinheiser HD 380 pro) and rated their overall perceived quality, as well as perceived amount of reverberation. To avoid biases, the order of presentation of the algorithms and conditions were randomized between subjects. Details on the listening test can be found in [11].

3.2. Figures-of-merit

Two figures-of-merit for instrumental measures are used, namely Pearson correlation coefficient between true and estimated quality (and perceived amount of reverberation) ratings, as well as the root-mean-square error (RMSE). Given the differences in anchors and hidden references for each condition, it is imperative that the performance of the measures be made for each acoustic condition separately. As such, correlation is compared for each of the four acoustic conditions. For overall comparison, the average correlation across all four conditions and the overall RMSE values are also reported.

4. EXPERIMENTAL RESULTS AND DISCUSSION

Table 2 reports the performances attained with the eleven instrumental measures as correlates of overall quality for each of the four acoustic conditions, as well as averaged over all four. Within the

Table 2: Performance comparison of instrumental measures as correlates of perceived quality

Metrics	S2,Near	S2,Far	R1, Near	R1, Far	AVG	RMSE
Non-intrusive instrumental measures						
SRMR	0.19	0.43	0.90	0.92	0.61	0.32
SRMR$_{\mathrm{norm}}$	**0.93**	**0.93**	**0.94**	**0.96**	**0.94**	**0.20**
ANIQUE+	0.87	0.41	0.57	0.05	0.47	0.26
P.563	0.12	0.51	0.67	0.61	0.47	0.31
Intrusive instrumental measures						
POLQA	0.56	0.75	0.76	0.78	0.71	0.24
PESQ	0.72	0.80	0.77	0.78	0.76	**0.21**
NCM	0.76	0.81	**0.91**	**0.91**	**0.84**	0.26
CSII	0.64	0.75	0.78	0.78	0.73	0.27
STOI	0.54	0.70	0.68	0.72	0.66	0.35
IS	0.81	**0.82**	0.74	0.68	0.76	**0.54**
LLR	**0.82**	**0.82**	0.74	**0.91**	0.82	0.51

Table 3: Performance comparison of instrumental measures as correlates of perceived amount of reverberation

Metrics	S2,Near	S2,Far	R1, Near	R1, Far	AVG	RMSE
Non-intrusive instrumental measures						
SRMR	0.00	0.53	0.75	0.76	0.72	0.37
SRMR$_{\mathrm{norm}}$	**0.86**	**0.84**	**0.88**	**0.96**	**0.88**	**0.22**
ANIQUE+	0.78	0.41	0.67	0.05	0.47	0.25
P.563	0.31	0.51	0.26	0.15	0.30	0.34
Intrusive instrumental measures						
POLQA	0.77	0.81	0.89	0.92	0.84	0.26
PESQ	**0.89**	**0.91**	**0.95**	**0.97**	**0.93**	**0.20**
NCM	0.87	0.79	0.87	0.89	0.85	0.26
CSII	0.83	0.77	0.84	0.87	0.82	0.25
STOI	0.72	0.69	0.74	0.80	0.73	0.32
IS	0.70	0.45	0.36	0.31	0.45	0.54
LLR	0.70	0.45	0.35	0.27	0.44	0.50

non-intrusive measures, SRMR achieved reliable performance at higher reverberation levels (i.e., "R1, Near" and "R1, Far"), but poor accuracy at lower reverberation levels. This is in line with the insights presented in [15, 37] and the higher RMSE shows the effects of speech content and pitch on the performance [21]. The extended version, SRMR$_{\mathrm{norm}}$, on the other hand, overcomes these limitations and results in a stable accuracy across all four tested acoustic conditions, as well as reduced RMSE by approximately 38%. The other two standard algorithms provided the same average correlation, with ANIQUE+ achieving lower RMSE values compared to ITU-T Rec. P.563. Notwithstanding, ANIQUE+, showed accurate correlation with MUSHRA scores in lower reverberation levels. Interestingly, the opposite was observed with P.563. It is hypothesized that the ANIQUE+ articulation analysis module performed better at lower reverberation levels, whereas the P.563 noise analysis module performed better at higher reverberation levels and these disturbances received higher weights during cognitive mapping. Overall, SRMR$_{\mathrm{norm}}$ achieved the highest correlation and lowest RMSE of all tested non-intrusive measures.

Regarding the intrusive measures, it can be seen that NCM resulted in the highest average correlation across the four acoustic conditions, with improved accuracy at higher reverberation levels. LLR and IS performed relatively stable across the tested conditions in terms of correlation, but achieved the highest RMSE values; overall, the LLR metric outperformed IS. As for the ITU-T standard metrics, POLQA and PESQ, interestingly their overall correlations were lower than the three abovementioned "classical" metrics. Overall, PESQ outperformed POLQA across the majority of the acoustic conditions and achieved somewhat lower RMSE values. This is an interesting finding, as POLQA is described as being applicable to enhanced speech, but the presented findings suggest otherwise. Lastly, the STOI and CSII measures showed similar behaviour across the four acoustic conditions, with improved accuracy at higher reverberation levels. Overall, the NCM metric showed the highest averaged correlation with the multi stimuli test with hidden reference and anchor (MUSHRA) scores and PESQ showed the lowest RMSE. Comparing the intrusive to non-intrusive metrics, it can be seen that SRMR$_{\mathrm{norm}}$ outperformed all other intrusive and non-intrusive measures in terms of correlation and achieved RMSE values in line with PESQ. This is an important finding, as SRMR$_{\mathrm{norm}}$ does not require access to a clean reference signal.

Results in Table 3 show the performances of the eleven instrumental measures as correlates of the perceived amount of reverberation. Similar to what was mentioned above, SRMR performance improved as reverberation levels increased and no correlation was seen in the 'S2, near condition'. SRMR$_{\mathrm{norm}}$, in turn, showed to be more stable across all conditions, outperforming all other considered non-intrusive measures in terms of both correlation and RMSE. As before, ANIQUE+ and P.563 achieved poor overall correlation, with ANIQUE+ achieving somewhat accurate correlation in the low reverberation level condition. Unlike the case of 2, however, P.563 performance did not improve with increasing reverberation levels, suggesting that an alternate mapping may be needed for the reverberation level rating scale. Overall, SRMR$_{\mathrm{norm}}$ achieved the highest correlation and lowest RMSE of all tested non-intrusive measures.

Regarding the ITU-T standard measures, POLQA and PESQ provided stable correlations across the four acoustic conditions, with PESQ outperforming POLQA by approximately 11% in terms of correlation. NCM and CSII achieved performance in line with POLQA, thus suggesting that further optimizations may be needed with POLQA in order for the measure to be reliably used with speech enhancement algorithms. As correlates of the amount of reverberation, both LPC-based metrics performed poorly and achieved the lowest correlation and highest RMSE of all tested instrumental measures. STOI performance was in line with the original non-intrusive SRMR metric. Overall, as correlates of perceived amount of reverberation, PESQ achieved the highest correlation and lowest RSME, followed closely by SRMR$_{\mathrm{norm}}$. Notwithstanding, as mentioned previously, the latter has the advantage of not requiring a clean reference signal.

5. CONCLUSION

In this paper, we compared the performance of eleven instrumental measures at assessing the quality and perceived amount of reverberation of speech processed by single- and multi-channel enhancement algorithms. Overall, it was observed that a recent extension to the non-intrusive SRMR metric, namely SRMR$_{\mathrm{norm}}$, outperformed all other metrics, including standard intrusive measures such as PESQ and POLQA. As correlates of perceived amount of reverberation, PESQ was shown to be the best instrumental measures, followed closely by SRMR$_{\mathrm{norm}}$ and POLQA. These results are promising, as SRMR$_{\mathrm{norm}}$ does not require access to a clean reference signal, thus is an ideal candidate for adaptive quality- or environment-aware speech enhancement.

6. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[2] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*, Springer Briefs in Electrical and Computer Engineering. Springer-Verlag, 2011.

[3] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sept. 2005.

[4] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383 –1393, May 2012.

[5] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech de-reverberation," *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[6] E. A. P. Habets, "Speech dereverberation using spectral enhancement," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds. Springer, 2010.

[7] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Multichannel signal enhancement algorithms for assisted listening devices," *IEEE Signal Processing Magazine*, vol. 32, pp. 18–30, Mar. 2015.

[8] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Multi-microphone noise reduction techniques as front-end devices for speech recognition," *Speech Communication*, vol. 34, pp. 3–12, 2001.

[9] Simon Doclo, Ann Spriet, Jan Wouters, and Marc Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction," *Speech Communication*, vol. 49, no. 7–8, pp. 636–656, Aug. 2007.

[10] S. Braun, D. P. Jarrett, J. Fischer, and E. A. P. Habets, "An informed spatial filter for dereverberation in the spherical harmonic domain," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 669–673.

[11] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, pp. 1–12, July 2015.

[12] K. Kinoshita et al., "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, pp. 1–19, Jan. 2015.

[13] Zenton Goh, Kah-Chye Tan, and T. G. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp. 287–292, May 1998.

[14] S. Moller, W.-Y. Chan, N. Cote, T. H. Falk, A. Raake, and M. Walermann, "Speech to: Models and trends," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 18–28, oct 2011.

[15] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Sept. 2010.

[16] Y. Hu and P. C. Loizou, "Objective measures for evaluating speech enhancement algorithms," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–237, 2008.

[17] S. Goetze et al., "A study on speech quality and speech intelligibility measures for quality assessment of single-channel dereverberation algorithms," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sept 2014, pp. 233–237.

[18] ITU-T, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms," Standard P.835, International Telecommunications Union (ITU-T), Nov. 2003.

[19] D. S. Kim and A. Tarraf, "Anique+: A new american national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Tech. J.*, vol. 12, pp. 221–236, 2007.

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

[21] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sept. 2014, pp. 55–59.

[22] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Feb. 2001.

[23] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2001, pp. 749–752.

[24] ITU-T, "Perceptual objective listening quality assessment: An advanced objective perceptual method for end-to-end listening speech quality evaluation of fixed, mobile, and IP-based networks and speech codecs covering narrowband, wideband, and super-wideband signals," Standard P.863, International Telecommunications Union (ITU-T), Jan. 2011.

[25] J. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part ii?perceptual model," *Audio Eng. Soc.*, vol. 61, no. 6, 2013.

[26] Jianfen Ma, Yi Hu, and Philipos C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3387–3405, May 2009.

[27] T. H. Falk, V. Parsa, J. F. Santos, K. A, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 32, no. 2, pp. 114–124, mar 2015.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4214–4217.

[29] James Kates and Kathryn Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 5, pp. 2224–2237, 2005.

[30] Jianfen Ma, Yi Hu, and Philipos C Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387–3405, 2009.

[31] Guo Chen, Soo Ngee Koh, and Ing Yann Soon, "Enhanced Itakura measure incorporating masking properties of human auditory system," *Signal Processing*, vol. 83, no. 7, pp. 1445 – 1456, 2003.

[32] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAMO: a british english speech corpus for large vocabulary continuous speech recognition," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Detroit, Michigan, U.S.A, May 1995, pp. 81–84.

[33] M. Lincoln, I. McCowan, J. Vepa, and H.K. Maganti, "The multichannel Wall Street Journal audio–visual corpus (MC-WSJ-AV): Specification and initial experiments," in *Proc. IEEE Workshop Autom. Speech Recognition and Understanding (ASRU)*, Cancún, Mexico, Dec. 2005, pp. 357–362.

[34] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[35] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, July 2001.

[36] ITU-T, "Method for the subjective assessment of intermediate quality levels of coding systems," Standard BS.1534–3, International Telecommunications Union (ITU-T), Nov. 2003.

[37] T. H. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 978–989, Apr. 2010.