A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech

Tiago H. Falk, *Member, IEEE*, Chenxi Zheng, and Wai-Yip Chan, *Member, IEEE*

***Abstract*—A modulation spectral representation is investigated for non-intrusive quality and intelligibility measurement of reverberant and dereverberated speech. The representation is obtained by means of an auditory-inspired filterbank analysis of critical-band temporal envelopes of the speech signal. Modulation spectral insights are used to develop an adaptive measure termed speech to reverberation modulation energy ratio. Experimental results show the proposed measure outperforming three standard algorithms for tasks involving estimation of multiple dimensions of perceived coloration, as well as quality measurement and intelligibility estimation of reverberant and dereverberated speech.**

***Index Terms*—Coloration, dereverberation, modulation spectrum, quality diagnosis, reverberation.**

I. Introduction

SPEECH acoustic signals propagating in enclosed environments are distorted by multiple reflections from the walls and other objects present in the room, hence making the speech signal sound "colored" and "reverberant" [1]. Coloration refers to the changes in signal timbre caused by early reflections [2], [3]. Late reflections, in turn, cause temporal smearing and the perceived effects depend on room geometry and wall sound absorption properties. Reverberation is known to degrade human-perceived speech quality and intelligibility as well as hamper automatic speech or speaker recognition performance. To compensate for such detrimental effects, dereverberation algorithms have been widely used. As emphasized in [4]; however, dereverberation is a difficult and often ill-conditioned problem, and can introduce objectionable artifacts to the processed speech signals. To evaluate the performance of dereverberation algorithms, subjective and/or objective quality and intelligibility measurement methods are needed.

Subjective methods require a listener panel to judge and quantify the quality and/or intelligibility of the processed speech signals. Commonly, subjective quality tests have listeners rate the quality of the speech signal on a pre-specified scale [5]. More recently, listening tests have also been used to characterize the subjective perception of coloration and reverberation decay tail effects [6]. Intelligibility, in turn, can be quantified using, for instance, nonsense syllable tests or consonant recognition tests wherein listeners mark on a test sheet the words or letters heard. Subjective tests are costly and labor-intensive, and perhaps more significantly, they are unsuitable for real-time applications. As a consequence, computer-based objective measurement methods have been the focus of more recent research efforts (e.g., [7]–[10]).

Objective measurement methods can be broadly classified as intrusive or non-intrusive. Intrusive measures depend on a distance metric between a clean reference speech signal and its reverberant or dereverberated counterpart. Non-intrusive measures, on the other hand, do not depend on a reference signal. To date, the majority of available standard objective *quality* measures have focused on transmission network-related distortions and have overlooked the effects of (de)reverberation on speech quality. Traditionally, conventional intrusive measures such as signal-to-noise ratio, bark spectral distortion [11], and cepstral distance [12] have been used. Such measures, however, have been shown to correlate poorly with subjective quality ratings [6]. In practice, original reference signals are seldom available. Reliable *non-intrusive* measures offer the flexibility needed to build practical real-time applications.

Objective *intelligibility* measures have been derived based on human perceptual concepts of temporal envelope modulations, making use of the so-called modulation transfer function [13]. The speech transmission index (STI) measure exemplifies the current state-of-the-art in objective intelligibility estimation [10]. While the standardized STI measure depends on artificial speech-like signals, several extensions have been proposed which allow for accurate estimation using the clean reference speech signal and its (de)reverberant counterpart in an *intrusive* manner [14]–[16]. To date, standardized non-intrusive intelligibility measurement methods are not available.

In this paper, perceptual insights are used to develop an adaptive *non-intrusive* measure termed speech-to-reverberation modulation energy ratio, based on extending the work described in [17]. More specifically, coloration and late reverberation effects are quantified in the modulation spectral domain and used to estimate 1) the quality components of a six-dimensional coloration space [18], 2) subjective scores of perceived reverberation tail effects and overall quality, as well as 3) intelligibility scores. Experiments suggest that the proposed measure outperforms three standardized quality measurement algorithms when estimating coloration, reverberation tail effects, and overall quality. Moreover, the proposed measure

Manuscript received September 25, 2009; revised February 01, 2010; accepted May 17, 2010. Date of current version August 13, 2010. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Biing-Hwang Juang .

T. H. Falk was with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON K7L 3N6, Canada. He is now with the Bloorview Research Institute/Bloorview Kids Rehab, University of Toronto, Toronto, ON M4G 1R8, Canada (e-mail: tiago.falk@ieee.org).

C. Zheng and W.-Y. Chan are with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: geoffrey.chan@queensu.ca).

Digital Object Identifier 10.1109/TASL.2010.2052247

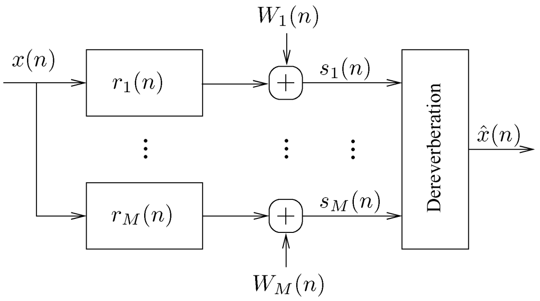


Fig. 1. Block diagram of multichannel speech dereverberation.

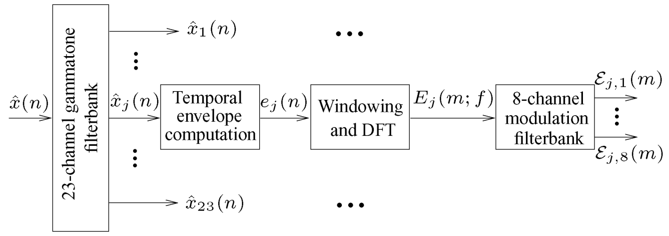


Fig. 2. Block diagram of the signal processing steps involved in the computation of modulation spectra.

attains performance comparable to a standardized intrusive method when estimating intelligibility scores, but adding the benefit of not requiring access to a reference signal.

The remainder of this paper is organized as follows. Section II presents a brief overview of multichannel dereverberation systems. Section III describes the signal processing and the motivation behind the proposed measure. Section IV reports experimental results along with database and benchmark algorithm descriptions. Conclusions are drawn in Section V.

II. Multichannel Dereverberation

Speech propagation from a speaker to a microphone placed in a reverberant room is conventionally modeled as a linear filtering process. In scenarios where M microphones are available, the reverberant signal $s_p(n)$, $p = 1, \ldots, M$ measured at the pth microphone is modeled as a convolution of the source (clean) speech signal $x(n)$ with the acoustic room impulse responses $r_p(n)$

$$s_p(n) = x(n) * r_p(n), \quad p = 1, \ldots, M. \tag{1}$$

If additive background noise $W_p(n)$ is present, (1) becomes

$$s_p(n) = x(n) * r_p(n) + W_p(n). \tag{2}$$

The ultimate goal in dereverberation is to derive a signal $\hat{x}(n)$ that is perceptually *imperceptible* from $x(n)$ by processing all the received signals $s_p(n), p = 1, \ldots, M$, as depicted in Fig. 1. In reality, since the room impulse responses $r_p(n)$ are unknown and time varying, dereverberation becomes a difficult blind estimation problem. Thus, dereverberation algorithms strive to *improve* the intelligibility of the reverberant signal while minimizing the introduction of unwanted artifacts, such as temporal discontinuities [4]. Dereverberation algorithms can be classified as single-microphone (or single-channel) or microphone array based (or multichannel), with the latter commonly providing improved performance [19].

In this paper, three conventional multichannel dereverberation paradigms are explored, namely, delay-and-sum beamforming, cepstral liftering, and blind subspace-based system identification (i.e., zero-forcing time domain dereveberation). A detailed description of the algorithms is beyond the scope of this paper and the reader is referred to [20] (and references therein) for more details regarding the algorithms and their associated parameters.

III. Modulation Spectral Signal Representation

The proposed measure is computed by performing spectral analysis on the modulation envelopes of the (de)reverberant speech signal. In this section, we first present the signal processing steps involved in the computation of our modulation spectral representation. The motivation for and the developed measure are then described.

A. Modulation Spectrum—Signal Processing

Fig. 2 depicts a block diagram of the signal processing steps used to compute our modulation spectral representation. Here, only a brief description is provided and the reader is referred to [21] for a more detailed explanation. First, the processed speech signal $\hat{x}(n)$ is filtered by a 23-channel gammatone filterbank to emulate the processing performed by the cochlea [22]. Filter center frequencies range from 125 Hz to nearly half the sampling rate; filter bandwidths are characterized by the equivalent rectangular bandwidth [23]. For simplicity, the remainder of this paper will use $\hat{x}(n)$ to denote the (de)reverberant speech signal.

The temporal envelope $e_j(n)$ of the j^{th} filter output signal $\hat{x}_j(n)$ is then computed using the Hilbert transform $\mathcal{H}\{\cdot\}$ as

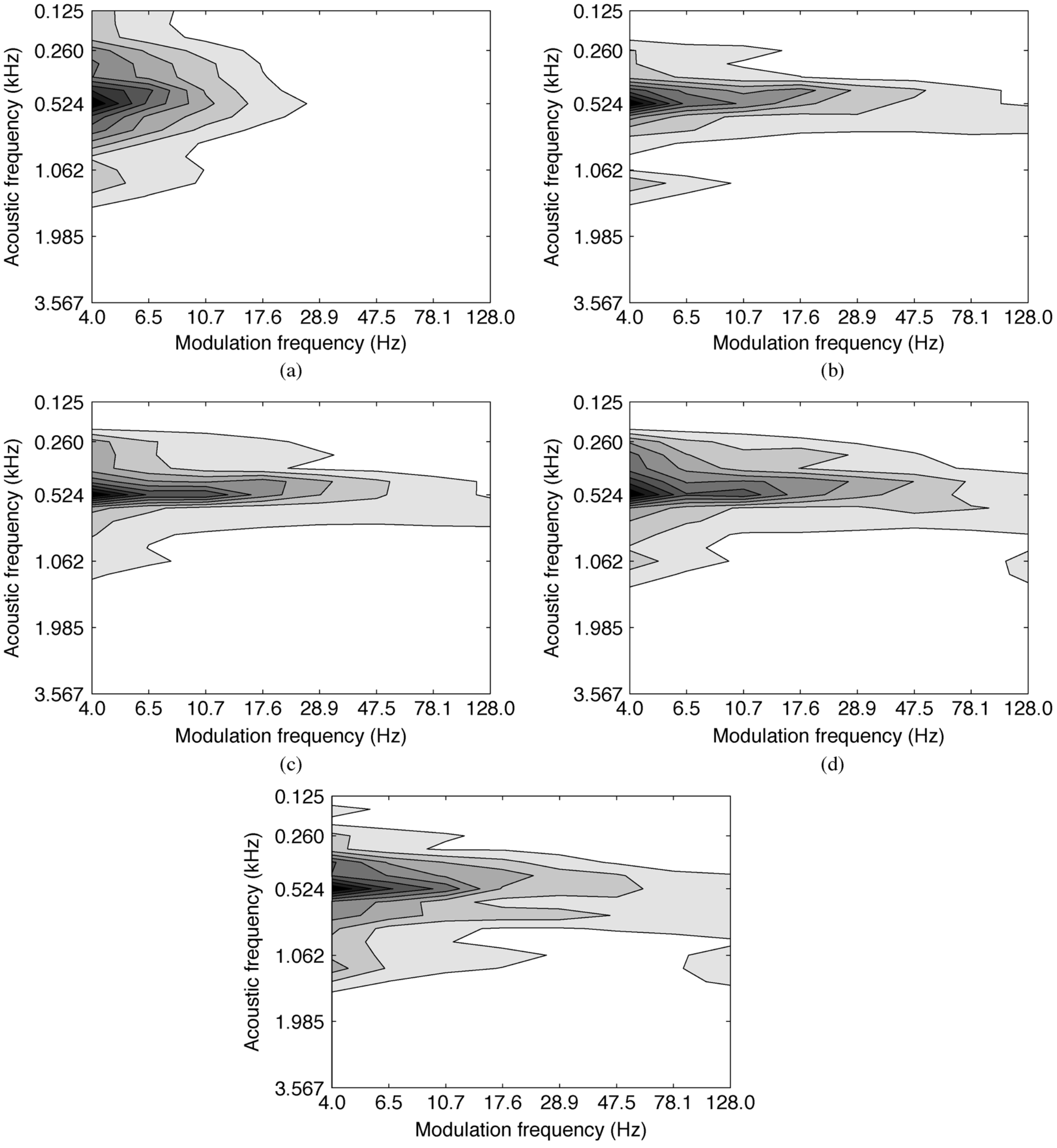
$$e_j(n) = \sqrt{\hat{x}_j(n)^2 + \mathcal{H}\left\{\hat{x}_j(n)\right\}^2}. \tag{3}$$

Temporal envelopes $e_j(n)$ are multiplied by a 256-ms Hamming window with 32-ms shifts and the windowed envelope for frame m is represented as $e_j(m;n)$. Frames of 256-ms duration are used in order to obtain appropriate resolution for low-frequency modulation frequencies around 4 Hz [24].

Modulation spectral energy for critical band j is then computed as the squared magnitude of the discrete Fourier transform $\mathcal{F}\{\cdot\}$ of the temporal envelope $e_j(m;n)$

$$E_j(m;f) = \left|\mathcal{F}\left(e_j(m;n)\right)\right|^2 \tag{4}$$

where f indexes the modulation frequency bins. Modulation frequency bins are grouped into eight bands in order to emulate an auditory-inspired modulation filterbank, as suggested by [25]. The notation $\bar{\mathcal{E}}_{j,k}$ is used to denote the average modulation energy over all frames of the j^{th} critical-band signal grouped by the k^{th} modulation filter, with $j = 1, \ldots, 23$, $k = 1, \ldots, 8$. Fig. 3(a) depicts a representative $\bar{\mathcal{E}}_{j,k}$ (also called modulation spectrogram) for a clean speech signal. The modulation spectrogram depicts the distribution of modulation energy as a function of modulation frequency and acoustic frequency, averaged



(a)

(b)

(c)

(d)

(e)

Fig. 3. Modulation spectrogram of (a) clean, (b) reverberant speech with $T_{60} = 533$ ms, and speech processed by (c) delay-and-sum beamformer, (d) cepstrum liftering, and (e) subspace-based dereverberation algorithms.

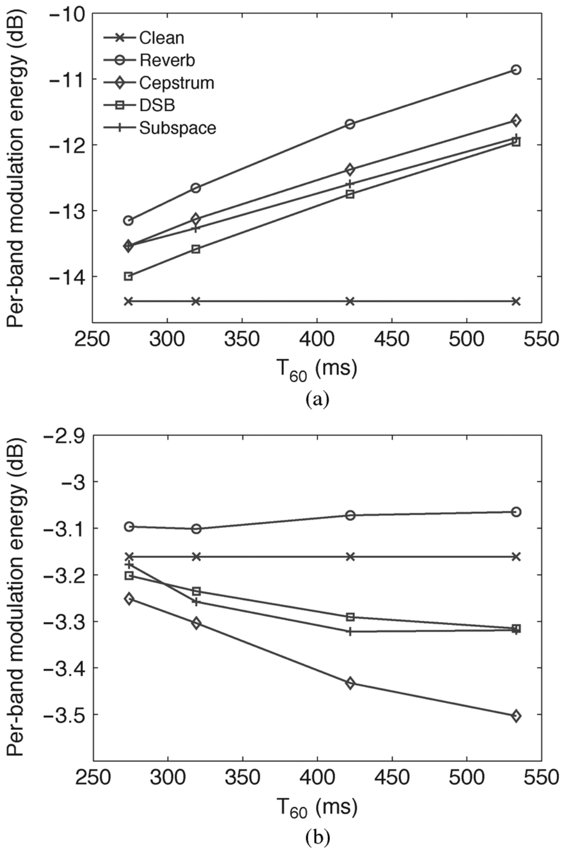
over all speech frames. Additionally, the average per-modulation band energy $\bar{\mathcal{E}}_k$ is denoted by

$$\bar{\mathcal{E}}_k = \frac{1}{23}\sum_{j=1}^{23}\bar{\mathcal{E}}_{j,k}. \tag{5}$$

B. Modulation Spectral Insights

Slow temporal envelope modulations have been shown to provide useful cues for objective quality [26] and intelligibility [27] estimation. It is known, for example, that for clean (anechoic) speech, temporal envelopes contain frequencies ranging from 2–20 Hz with spectral peaks at approximately 4 Hz, which corresponds to the syllabic rate of spoken speech [28] [see Fig. 3(a)]. With reverberant speech, the diffuse reverberation tail is often modeled as an exponentially damped Gaussian white noise process. With increasing reverberation levels, the signal attains more Gaussian white-noise like properties. Given the property that temporal envelopes, computed via a Hilbert transformation, can contain frequencies up to the bandwidth of the envelope bearing signal [30], it is expected that reverberant signals exhibit higher-frequency temporal envelopes due to the "whitening" effect of the reverberation tail [31].

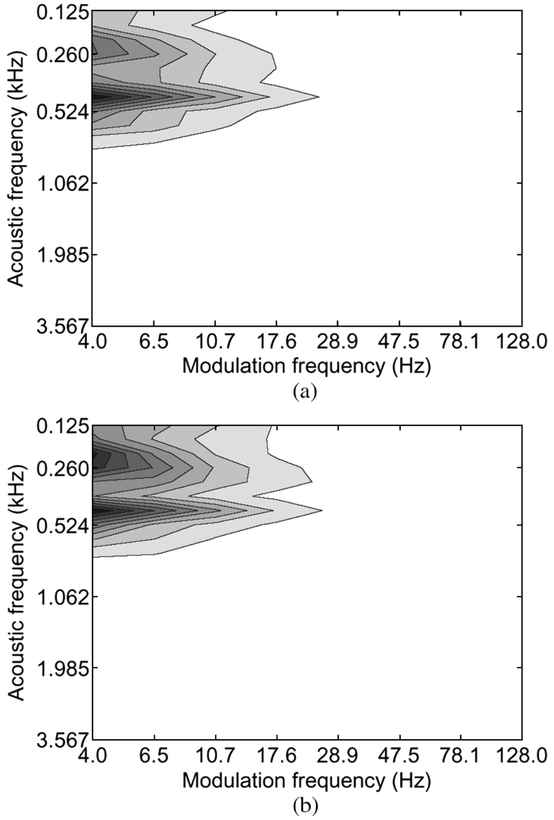
This property is illustrated with the modulation spectrograms depicted in Fig. 3. Subplots (a) and (b) illustrate the modulation spectrogram for a clean and reverberant speech signal with a reverberation time $T_{60} = 533$ ms, respectively. As can be seen, for clean speech, the bulk of the modulation energy is



(a)

(b)

Fig. 4. Per-band modulation energy versus reverberation time (T_{60}) for modulation band (a) $k = 1$ (∼4 Hz) (b) $k = 6$ (∼50 Hz).



(a)

(b)

Fig. 5. Modulation spectrogram of (a) clean speech signal and (b) its colored counterpart.

situated at below 20 Hz, and peaks at around 4 Hz. Reverberation, on the other hand, causes smearing of the energy into higher modulation frequencies. Subplots (c)–(e), in turn, depict modulation spectrograms of the reverberant speech signal after processing by a delay-and-sum beamformer (DSB), cepstral liftering and subspace-based dereverberation algorithms, respectively. As observed, high-frequency modulation energy is still present post dereverberation, thus suggesting lower quality and intelligibility relative to clean speech.

As such, suitably crafted features extracted from the modulation spectrum can provide useful information for non-intrusive quality and intelligibility measurement. To further investigate the effects of multichannel dereverberation on the modulation spectrum, 330 anechoic speech signals are convolved with room impulse responses measured by a linear microphone array in four different enclosures (reverberation time values of $T_{60} = 274$, 319, 422, and 533 ms) [20]. The three dereverberation algorithms described in Section II are applied to the reveberated signals. For the deverberation processed signals, Fig. 4(a) and (b) plots $\bar{\mathcal{E}}_k$ for modulation bands $k = 1$ and $k = 6$, corresponding to modulation frequencies around 4 Hz and 50 Hz, respectively.

As seen from Fig. 4(a), low-frequency modulation energy is slightly increased (∼0.1 dB) for reverberant speech relative to clean speech. The effect, however, is shown to be relatively independent of reverberation time and is likely due to early reflections. This conjecture is corroborated by the experiments with an artificially generated coloration dataset, reported in [18], and the illustration in Fig. 5. As can be seen, early reflections emphasize modulation frequency content around 4 Hz. As such, the early reflections likely cause the improved intelligibility that has been observed with strong early reflections whose delay times are around 50 ms [32]. Fig. 4(a) also shows that the dereverberation algorithms decrease the low-frequency modulation energy by between 0.1–0.3 dB below clean speech. The decrease is likely due to introduced artifacts which degrades intelligibility. At a reverberation time $T_{60} = 533$ ms, cepstral liftering suppresses low-frequency modulation content the most.

Fig. 4(b) shows the higher modulation frequency channel exhibiting a stronger dependency of modulation energy on reverberation time. The modulation energy (in dB) increases almost linearly with reverberation time. The delay-and-sum beamformer is shown to attain the most suppression and reduce the high-frequency modulation energy by approximately 1 dB relative to reverberant speech. The gain, however, is still modest; an approximately 2.5 dB difference remains between anechoic and dereverberated speech for reverberation time of 533 ms. Such difference is due to the residual reverberation tail that remains post dereverberation.

C. Proposed Measure

Using the insights described above, an adaptive measure termed speech to reverberation modulation energy ratio

(SRMR) is proposed for non-intrusive quality diagnosis of (de)reverberant speech. The measure is given by

$$\mathrm{SRMR} = \frac{\sum_{k=1}^{4} \bar{\mathcal{E}}_k}{\sum_{k=5}^{K^*} \bar{\mathcal{E}}_k} \qquad (6)$$

where the upper summation bound K^* in the denominator is adapted to the speech signal under test. As mentioned in Section III-B, modulation frequency content for acoustic frequency band j is upper-bounded by the bandwidth of critical-band filter j. As such, speech signals with different acoustic frequency content, subjected to the same reverberation effects, can result in different modulation spectra.

Plots in Fig. 6(a) and (b) illustrate a representative example where the percentage of modulation energy present per acoustic frequency channel is plotted versus acoustic frequency. The plots are for anechoic speech produced by two different speakers and then convolved with a room impulse response with a reverberation time of $T_{60} = 319$ ms. As can be seen, for subplot (a), 90% of the total modulation energy is obtained below 600 Hz. For subplot (b), in turn, 90% of the total energy is obtained below 1 kHz. The bandwidths of the gammatone filters centered at such frequencies are 86 Hz and 131 Hz, respectively. As a consequence, due to properties of the modulation filterbank [21], negligible energy at modulation frequency band $k = 8$ (centred at around 128 Hz) is expected from the signal represented in subplot (a). In the experiments described in Section IV, K^* is chosen on a per-signal basis and depends on the bandwidth of the lowest gammatone filter for which 90% of the total energy is accounted for. As examples, for the speech signals represented in Fig. 6(a) and (b), $K^* = 7$ and $K^* = 8$, would be used, respectively.

IV. Experimental Results

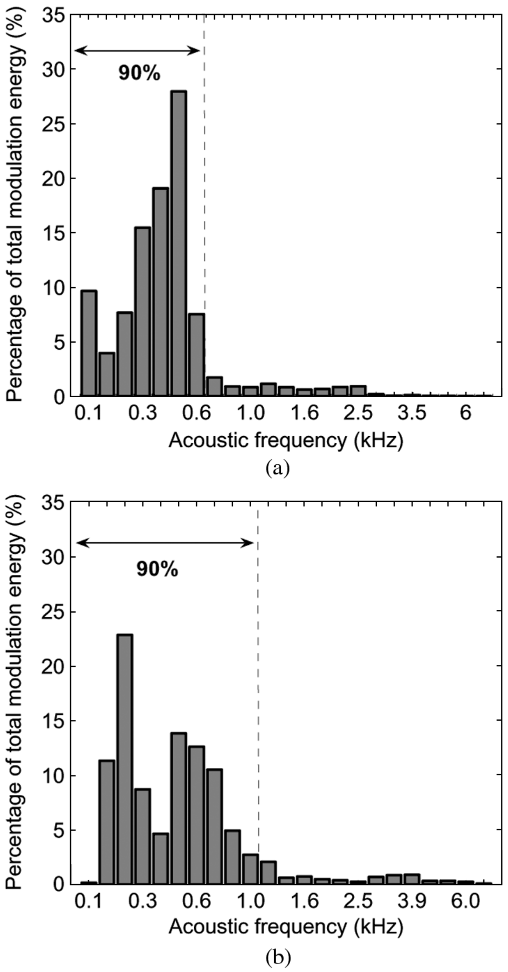
In this section, the three datasets used in the experiments are described, benchmark algorithms are detailed, and quality and intelligibility estimation results are presented.

A. Database Description

Three databases are used in our experiments and are detailed in the subsections below.

1) Database 1—Multidimensional Coloration Space: The first database is used to investigate the effectiveness of the proposed measure in estimating multiple dimensions of perceived coloration. Different coloration effects are artificially generated by manipulating three coloration control parameters, namely, spectral roughness, spectral tilt, and local spectral extremes. Speech signals are digitized with 16-bit precision and 22.5-kHz sampling rate. The reader is referred to [18] for more details.

A subjective verbal attribute listening test was performed with 16 expert listeners (with audio or musical background), all male with no reported hearing loss. Subjects were presented with the reference anechoic speech signal and its colored counterpart and were asked to rate the latter using six attributes: *warm*, *thin*, *cold*, *bright*, *boomy*, and *muffled*. Following the suggestions of [33], each attribute is rated on a nine-point scale that is anchored by the attribute at one end and its opposite at the other end, e.g., thin and not thin. Seventeen different coloration-distorted speech files were generated from each clean speech file, comprised of a concatenated female- and male-uttered sentence to minimize the bias of speaker dependent characteristics [18]. The subjective ratings for each attribute were averaged over all the listeners to create six mean opinion scores for each speech file.



(a)

(b)

Fig. 6. Percentage of total modulation energy, per acoustic frequency band, for speech signals from two different speakers.

2) Database 2—(De) Reverberation Quality: The second database is a subjectively scored multichannel acoustic reverberation database termed Multichannel Acoustic Reverberation Database at York (MARDY) developed for evaluation of dereverberation algorithms [6]. The database uses room impulse responses which were collected with a linear microphone array in an anechoic chamber with reflective panels and absorptive panels in place. Speaker to microphone distances varied between 1–4 m (1-m increments) and reverberation time values ranged from $T_{60} = 291$ ms to 447 ms. Reverberant speech was generated with the collected room impulse responses and anechoic speech from two speakers (one male and one female).

Three dereverberation paradigms were tested, namely, delay-and-sum beamforming, a proprietary multichannel method based on a statistical model of late reverberation and spectral subtraction, and a proprietary multi-microphone method based on spatio-temporal averaging operating on the linear prediction residual; the reader is referred to [6] for more details. The positions of the source and microphones were assumed to be known for all three methods. As the proprietary portion of the database is not publicly available, for the experiments described herein, only the reverberant speech signals and the signals processed by the delay-and-sum beamformer are used. Speech signals are digitized with 16-bit precision and 16-kHz sampling rate.

A multidimensional subjective listening test was performed following the guidelines of the International Telecommunications Union ITU-T Recommendation P.835 [34]. In the test, 26 normal hearing listeners rated the subjective perception of coloration (COL), reverberation tail effect (RTE), and overall speech quality (MOS) for 32 speech signals uttered by both male and female speakers. For each category, listeners used a 5-point scale where a rating of 5 indicated the best score and a rating of 1 the worst score. Speech examples were presented to the listeners in order to familiarize them with identification and quantification of coloration and reverberation tail effects.

3) Database 3—(De) Reverberation Intelligibility: The third database consists of a modified version of the popular Wall Street Journal November 92 speech recognition evaluation test set. The original dataset consists of 330 sentences uttered by eight different speakers, both male and female, in clean conditions. The modified version consists of the 330 aforementioned speech signals convolved with six-channel room impulse responses measured by a linear microphone array in four different enclosures with reverberation times of T_{60} =274, 319, 422, and 533 ms [20]. Reverberant speech signals are further processed by the three dereverberation algorithms described in Section II. Speech signals are digitized with 16-bit precision and 16-kHz sampling rate.

Motivated by the work described in [35], three speech-based derivatives of the popular speech transmission index (STI) are used as measures of speech intelligibility. The three intrusive measures were proposed by Payton [15], [36], Drullman [14], [37], and Goldsworthy [16]; a detailed description of the signal processing computation for the three measures is given in [16]. Previous research has suggested that the three measures are reliable predictors of speech intelligibility for nonlinear distortion conditions such as (de)reverberation [38], [39], with the method proposed by Goldsworthy attaining superior performance [16].

B. Benchmark Algorithms

The performance of the proposed SRMR measure is compared to that of three standard quality measurement algorithms, two of which are non-intrusive. The intrusive algorithm is the ITU-T standard P.862 algorithm, better known as Perceptual Evaluation of Speech Quality (PESQ) which has a narrowband (8-kHz sample rate) [7] and a wideband (16 kHz) [40] version. With PESQ, both the reference and processed (reverberant or dereverberated) signals are transformed to a psychophysical representation by means of perceptual frequency mapping and compressive loudness scaling. The difference between the psychophysical representations of the degraded and reference speech signals is then calculated and mapped to a quality score using a cognitive-like regression model. PESQ has been widely used for quality measurement of network transmitted speech and represents the current state-of-the-art in intrusive quality measurement. Its use, however, is not recommended for reverberant or dereverberated speech [7], [41]; nonetheless, recent research has suggested accurate ratings for reverberant speech [42].

The two non-intrusive standard measures include the ITU-T standard P.563 [8] and the American National Standards Institute ANSI standard ANIQUE+ [9]. The P.563 algorithm combines three principles for speech quality measurement [43]. First, vocal tract and linear prediction analysis is performed to detect unnaturalness in the speech signal. Second, a pseudo-reference signal is reconstructed by modifying the computed linear prediction coefficients to fit the vocal tract model of a typical human speaker. The pseudo-reference signal serves as input, along with the degraded speech signal, to a double-ended algorithm (similar to ITU-T P.862) to generate a basic voice quality measure. Lastly, specific distortions such as noise, temporal clippings, and robotization effects (voice with metallic sounds) are detected.

A total of 51 characteristic signal parameters are calculated and based on a restricted set of eight key parameters, one of six major distortion classes is detected. The distortion classes are, in decreasing order of annoyance: high level of background noise, signal interruptions, signal-correlated noise, speech robotization, and unnatural male and female speech [43]. For each distortion class, a subset of the extracted parameters is used to compute an intermediate quality rating. Once a major distortion class is detected, its intermediate score is linearly combined with eleven other parameters to derive a final quality estimate. P.563 represents the current state-of-the-art in non-intrusive quality measurement. While the algorithm has demonstrated acceptable accuracy for transmission systems with echo cancelers [8], recent research has reported poor correlation with subjective quality ratings for reverberant and dereverberated speech [17], [42].

The second non-intrusive benchmark algorithm is ANIQUE+. The algorithm became an ANSI standard after being “narrowly beaten” by P.563 in the ITU-T competition to standardize a non-intrusive model in 2004 [44]. The algorithm is based on three distortion measurement modules: mute, non-speech, and articulation. The mute distortion module detects unnatural mutes in the speech signals and quantifies their effects on speech quality. The non-speech module, in turn, detects and quantifies the effects of annoying non-speech activities, such as those resultant from inserting erroneous bits into a speech decoder [45]. Lastly, the articulation distortion module uses modulation spectral concepts similar to those used in the proposed measure. More specifically, ANIQUE+ computes for each critical band a so-called normalized articulation energy (average modulation energy between 2–30 Hz modulation frequencies), normalized non-articulation energy (average modulation energy for frequencies greater than 30 Hz), and the energy across the critical band. The three entities computed for all the critical bands are mapped to a frame distortion score by means of a multilayer perceptron. The frame distortion scores are aggregated, separately over active and inactive frames. The

TABLE I

PERFORMANCE COMPARISON BETWEEN SRMR, PESQ, P.563, AND ANIQUE+ ON DATABASE 1. COLUMN LABELED "% ↑" INDICATES THE CORRELATION IMPROVEMENT GIVEN BY (7). AVERAGE CORRELATION IMPROVEMENT IS COMPUTED OVER THE THREE BENCHMARK ALGORITHMS

Algorithm	Warm	% ↑	Thin	% ↑	Cold	% ↑	Bright	% ↑	Boomy	% ↑	Muffled	% ↑
SRMR	0.89	–	-0.86	–	-0.88	–	-0.82	–	0.86	–	0.78	–
PESQ	0.60	72.5	-0.62	63.2	-0.58	71.4	-0.38	71.0	0.57	67.4	0.33	67.2
P.563	0.43	80.7	-0.47	73.6	-0.48	76.9	-0.18	78.0	0.48	73.1	0.18	73.2
ANIQUE+	0.57	74.4	-0.57	67.4	-0.54	73.9	-0.17	78.3	0.56	68.2	0.15	74.1
Average	–	75.9	–	68.1	–	74.1	–	75.8	–	69.6	–	71.5

TABLE II

PERFORMANCE COMPARISON BETWEEN SRMR, PESQ, P.563, AND ANIQUE+ ON DATABASE 2. COLUMN LABELED "% ↑" INDICATES THE CORRELATION IMPROVEMENT GIVEN BY (7). AVERAGE CORRELATION IMPROVEMENT IS COMPUTED OVER THE THREE BENCHMARK ALGORITHMS

	Overall (reverberant + dereverberated)						Reverberant						Delay-and-sum					
Algorithm	COL	%↑	RTE	%↑	MOS	%↑	COL	%↑	RTE	%↑	MOS	%↑	COL	%↑	RTE	%↑	MOS	%↑
SRMR	0.84	–	0.82	–	0.79	–	0.81	–	0.82	–	0.79	–	0.87	–	0.83	–	0.80	–
PESQ	0.66	52.9	0.81	5.3	0.72	25.0	0.66	44.1	0.81	5.3	0.70	30.0	0.67	60.6	0.82	5.6	0.78	9.1
P.563	0.44	71.4	0.46	66.7	0.35	67.7	0.38	69.4	0.41	69.5	0.31	69.6	0.54	71.7	0.50	66.0	0.40	66.7
ANIQUE+	0.72	42.9	0.70	40.0	0.77	8.7	0.77	17.4	0.76	25.0	0.84	-31.3	0.67	60.6	0.57	60.5	0.67	39.4
Average	–	55.7	–	37.3	–	33.8	–	43.6	–	33.3	–	22.8	–	64.3	–	44.0	–	38.4

three distortion modules' outputs are finally linearly combined to produce an overall quality score.

C. Multidimensional Coloration Estimation Performance

Table I reports correlation values (ρ) attained between the proposed measure and the multidimensional subjective coloration ratings available with Database 1 (see Section IV-A1); performance is compared to that obtained with the three benchmark algorithms. Since the majority of the benchmark algorithms operate at an 8-kHz sampling rate, results reported throughout the remainder of this paper will be based on subsampled versions of the databases described in Section IV-A. The column labeled "$\% \uparrow$" lists the percentage improvement in correlation obtained by using SRMR relative to algorithm "X." The correlation improvement is computed as

$$\% \uparrow = \frac{|\rho_{SRMR}| - |\rho_X|}{1 - |\rho_X|} \times 100\% \tag{7}$$

and indicates percentage reduction of the performance gap of algorithm "X" to perfect correlation. Note that the correlation signs in Table I are consistent with Table 1 in [18]. As can be seen, the proposed measure outperforms the three benchmark algorithms for all six dimensions in the coloration space. Correlation improvements, averaged over the benchmark algorithms, are greater than 68% for all dimensions, with average improvements of up to 75.9% being observed for dimension "warm." Performance improvements are more pronounced relative to the two benchmark non-intrusive algorithms.

D. Quality Measurement Performance

Table II reports correlation values attained between the three subjective scores available with Database 2 and the proposed measure and three benchmark algorithms. As observed, the proposed measure is shown to reliably estimate the three quality dimensions for both reverberant and dereverberated speech. Overall, SRMR is shown to outperform the intrusive and non-intrusive benchmark algorithms by an average 55%, 37%, and 33% for the COL, RTE, and MOS dimensions, respectively. For dereverberated speech, higher gains are observed and average improvements of 64%, 44%, and 38% are attained for the COL, RTE, and MOS dimensions, respectively. For reverberant speech, ANIQUE+ is shown to outperform SRMR in MOS estimation. Notwithstanding, the capability of the proposed measure to reliably estimate coloration and reverberation tail effects, in addition to overall quality, suggests it is a more suitable candidate for non-intrusive evaluation of reverberant speech and dereverberation algorithms, such as the delay-and-sum beamformer.

E. Intelligibility Estimation Accuracy

Table III reports correlation values attained between the three STI measures computed for Database 3 and the proposed measure and three benchmark algorithms. The columns labeled "STI_i," i =1–3, correspond to the STI measures computed by the intrusive methods described in [14]–[16], respectively. The "Reverberation" condition refers to the reverberant speech signal captured by the third microphone in the microphone array. As observed, the proposed SRMR measure attains higher correlation with STI_3, thus corroborates findings reported in [16] that STI_3 is more reliable for reverberant speech. Focusing on STI_3, the proposed measure is shown to improve over PESQ, P.563, and ANIQUE+ by an average 33.5%, 92.4%, and 89%, respectively. The high correlations reported by PESQ corroborate those reported in [46]. The proposed measure, however, allows for reliable intelligibility estimation *without* needing a reference signal.

V. CONCLUSION

A speech to reverberation modulation energy ratio measure is proposed for *non-intrusive* quality and intelligibility estima-

TABLE III
CORRELATION BETWEEN SRMR, PESQ, P.563, OR ANIQUE+ AND STI VALUES COMPUTED BY THE INTRUSIVE METHODS OF DRULLMAN [14] (STI_1), PAYTON [15] (STI_2), AND GOLDSWORTHY [16] (STI_3) USING DATABASE 3. COLUMN LABELED "% ↑" INDICATES THE CORRELATION IMPROVEMENT OVER STI_3 AS GIVEN BY (7). AVERAGE CORRELATION IMPROVEMENT IS COMPUTED OVER THE FOUR DEGRADATION CONDITIONS

	SRMR			PESQ				P.563				ANIQUE+			
Algorithm	STI_1	STI_2	STI_3	STI_1	STI_2	STI_3	% ↑	STI_1	STI_2	STI_3	% ↑	STI_1	STI_2	STI_3	% ↑
Reverberation	0.92	0.94	0.96	0.88	0.92	0.92	50.0	0.10	0.11	0.10	95.6	0.42	0.43	0.41	93.2
DSB	0.90	0.92	0.96	0.89	0.93	0.94	33.3	0.12	0.12	0.11	95.5	0.45	0.45	0.46	92.6
Cepstrum	0.90	0.93	0.95	0.86	0.91	0.92	37.5	0.06	0.07	0.06	94.7	0.47	0.48	0.50	90.0
Subspace	0.81	0.86	0.87	0.78	0.85	0.85	13.3	0.20	0.19	0.19	84.0	0.28	0.30	0.34	80.3
Average	–	–	–	–	–	–	33.5	–	–	–	92.4	–	–	–	89.0

tion of reverberant and dereverberated speech. The performance of the proposed measure is compared to that of three standard measurement algorithms, namely, ITU-T PESQ, ITU-T P.563, and ANSI ANIQUE+, using three databases. The first database is used to explore the performance of the algorithms in estimating multiple dimensions of perceived coloration. The second and third databases are used to investigate quality measurement and intelligibility estimation performance, respectively. Experimental results show the proposed measure outperforming all three standard algorithms on all three experiments. A Matlab implementation of the proposed measure can be made available for research purposes by contacting the first author.

ACKNOWLEDGMENT

The authors would like to thank Dr. K. Eneman for providing the dereverberation algorithms and multichannel room impulse responses and Dr. J. Wen for making the MARDY and the multidimensional coloration databases available.

REFERENCES

[1] D. Berkley, "Normal listeners in typical rooms—Reverberation perception, simulation, and reduction," in *Acoustical Factors Affecting Hearing Aid Performance*. Baltimore, MD: University Park Press, 1980, pp. 3–24.

[2] T. Halmrast, "Sound coloration from (very) early reflections," in *Proc. Meeting Acoust. Soc. Amer.*, Jun. 2001.

[3] P. Rubak, "Coloration in room impulse responses," in *Proc. Joint Baltic-Nordic Acoust. Meeting*, Jun. 2004, pp. 1–14.

[4] Y. Huang, J. Benesty, and J. Chen, "Speech enhancement: Dereverberation," in *Handbook of Speech Processing*. New York: Springer, 2008, pp. 929–943.

[5] ITU-T P.800, "Methods for subjective determination of transmission quality," Int. Telecom. Union, 1996.

[6] J. Wen, N. Gaubitch, E. Habets, T. Myatt, and P. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2006.

[7] ITU-T P.862, "Perceptual evaluation of speech quality: An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Int. Telecom. Union, 2001.

[8] ITU-T P.563, Single-ended method for objective speech quality assessment in narrowband telephony applications Int. Telecom. Union, 2004.

[9] ATIS-PP-0100005.2006, Auditory non-intrusive quality estimation plus (ANIQUE+): Perceptual model for non-intrusive estimation of narrowband speech quality Amer. National Standards Inst., 2006.

[10] BS EN 60268-16:2003, Sound system equipment. Objective rating of speech intelligibility by speech transmission index British Standards Inst., 2003.

[11] S. Wang, A. Sekey, A. Gersho, T. Syst, and C. Berkeley, "An objective measure for predicting subjective quality of speechcoders," *IEEE J. Sel. Areas Commun.*, vol. 10, no. 5, pp. 819–829, Jun. 1992.

[12] A. Gray, Jr. and J. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 5, pp. 380–391, Oct. 1976.

[13] H. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Amer.*, vol. 67, p. 318, 1980.

[14] R. Drullman, J. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 5, pp. 2670–2680, May 1994.

[15] K. Payton and L. Braida, "A method to determine the speech transmission index from speech waveforms," *J. Acoust. Soc. Amer.*, vol. 106, p. 3637, 1999.

[16] R. Goldsworthy and J. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Amer.*, vol. 116, p. 3679, 2004.

[17] T. H. Falk and W.-Y. Chan, "A non-intrusive quality measure of dereverberated speech," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Sep. 2008.

[18] J. Wen and P. Naylor, "Semantic coloration space investigation: Controlled coloration in the bark-sone domain," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 2007, pp. 311–314.

[19] *Speech and Audio Processing in Adverse Environments*, E. Hansler and G. Schmidt, Eds. New York: Springer, 2008.

[20] K. Eneman and M. Moonen, "Multimicrophone speech dereverberation: Experimental validation," *EURASIP J. Audio, Speech, Music Process.*, 2007, 19 pages.

[21] T. H. Falk and W.-Y. Chan, "Modulation spectral features for robust far-field speaker identification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 90–100, Jan. 2009.

[22] M. Slaney, "An Efficient implementation of the Patterson–Holdsworth auditory filterbank," Apple Computer, 1993, Tech. Rep..

[23] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, no. 1, pp. 103–138, 1990.

[24] R. Drullman, J. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1053–1064, Feb. 1994.

[25] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. I—Model structure," *J. Acoust. Soc. Amer.*, vol. 99, no. 6, pp. 3615–3622, 1996.

[26] D.-S. Kim, "A cue for objective speech quality estimation in temporal envelope representation," *IEEE Signal Process. Lett.*, vol. 11, no. 10, pp. 849–852, Oct. 2004.

[27] T. Houtgast and H. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Amer.*, vol. 77, no. 3, pp. 1069–1077, Mar. 1985.

[28] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes," in *Proc. Int. Conf. Speech Lang. Process.*, Oct. 1996, pp. 2490–2493.

[29] D.-S. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 821–831, Sep. 2005.

[30] Z. Smith, B. Delgutte, and A. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Lett. Nature*, vol. 416, pp. 87–90, Mar. 2002.

[31] 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.

[32] Y. Oh, D. Jeong, S. Doo, H. Lee, C. Choi, L. Kim, and I. Ko, "Spatial distribution of early reflections and speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 109, pp. 2313–2314, 2001.

[33] J. Holt, "Sounds like? An audio glossary," *Stereophile Mag.*, vol. 16, no. 7, pp. 1–16, Jul. 1993.

[34] ITU-T P.835, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms," Int. Telecom. Union, 2003.

[35] K. Paliwal, K. Wojcicki, and K. Wheeler, "Effect of analysis window duration on speech intelligibility," *IEEE Signal Process. Lett.*, vol. 15, pp. 785–788, 2008.

[36] K. Payton, L. Braida, S. Chen, P. Rosengard, and R. Goldsworthy, "Computing the STI using speech as a probe stimulus," in *Past, Present, and Future of the Speech Transmission Index*. Soesterberg, The Netherlands: TNO Human Factors, 2002, pp. 125–138.

[37] R. Drullman, "Temporal envelope and fine structure cues for speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 97, p. 585, 1995.

[38] H. Steeneken and T. Houtgast, "Validation of the revised STI method," *Speech Commun.*, vol. 38, no. 3–4, pp. 413–425, 2002.

[39] S. Tang and M. Yeung, "Reverberation times and speech transmission indices in classrooms," *J. Sound Vibr.*, vol. 294, no. 3, pp. 596–607, 2006.

[40] ITU-T P.862.2, "Wideband extension to Rec. P.862 for the assessment of wideband telephone networks and speech codecs," Int. Telecom. Union, 2007.

[41] ITU-T P.862.3, "Application guide for objective quality measurement based on recommendations P.862, P.862.1 and P.862.2," Int. Telecom. Union, 2005.

[42] A. de Lima, F. Freeland, P. Esquef, L. Biscainho, B. Bispo, R. de Jesus, S. Netto, R. Schafer, A. Said, B. Lee, and A. Kalker, "Reverberation assessment in audioband speech signals for telepresence systems," in *Proc. Int. Conf. Signal Process. Multimedia Applicat.*, Jul. 2008, pp. 257–262.

[43] 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, Nov. 2006.

[44] A. Rix, J. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza, "Objective assessment of speech and audio quality—Technology and applications," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1890–1901, Nov. 2006.

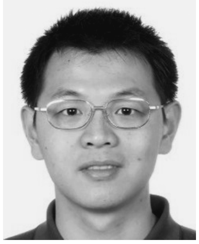
[45] D.-S. Kim and A. Tarraf, "ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Tech. J.*, vol. 12, no. 1, pp. 221–236, May 2007.

[46] J. Beerends, E. Larsen, N. Lyer, and J. van Vugt, "Measurement of speech intelligibility based on the PESQ approach," in *Proc. Int. Conf. Meas. Speech Audio Quality Netw.*, 2004, 4 pages.

Tiago H. Falk (S'00–M'09) received the B.Sc. degree from the Federal University of Pernambuco, Recife, Brazil, in 2002, and the M.Sc. and Ph.D. degrees from Queen's University, Kingston, ON, Canada, in 2005 and 2008, respectively, all in electrical engineering.

He is currently a Postdoctoral Fellow at the Bloorview Research Institute, affiliated with the University of Toronto, Toronto, ON, Canada. His research interests include multimedia quality measurement and enhancement, biomedical signal processing, rehabilitation engineering, and assistive technology development.

Dr. Falk is recipient of the IEEE Kingston Section Ph.D. Research Excellence Award (2008), the Best Student Paper Awards at ICASSP (2005) and IWAENC (2008), and the Newton Maia Young Scientist Award (2001).



Chenxi Zheng received the B.Sc degree in electrical engineering from Nanjing University of Science and Technology, Nanjing, China. He is currently pursuing the M.Sc. degree at Queen's University, Kingston, ON, Canada.

His research interests include speech enhancement and speech quality measurement.

Wai-Yip Chan (M'02) received the B.Eng. and M.Eng. degrees from Carleton University, Ottawa, ON, Canada, and the Ph.D. degree from the University of California, Santa Barbara, all in electrical engineering.

He is currently with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada. He has held positions with the Communications Research Centre, Bell Northern Research (Nortel), McGill University, and the Illinois Institute of Technology. His research interests are in multimedia signal processing and communications. He is an Associate Editor of the *EURASIP Journal on Audio, Speech, and Music Processing*.

Dr. Chan is a member of the IEEE Signal Processing Society Speech and Language Technical Committee. He has helped organize IEEE-sponsored conferences on speech coding, image processing, and communications. He received a CAREER Award from the U.S. National Science Foundation.