



On the use of the i-vector speech representation for instrumental quality measurement

Anderson R. Avila^{1,2} · Jahangir Alam² · Douglas O'Shaughnessy¹ · Tiago H. Falk¹

Received: 7 November 2019 / Published online: 20 June 2020
© Springer Nature Switzerland AG 2020

Abstract

The i-vector framework has been widely used to summarize speaker-dependent information present in a speech signal. Considered the state-of-the-art in speaker verification for many years, its potential to estimate speech recording distortion/quality has been overlooked. This paper is an attempt to fill this gap. We conduct a detailed analysis of how distortions are captured in the total variability space. We then propose a full-reference speech quality model based on i-vector similarities and three no-reference approaches. The first no-reference approach makes use of a single reference i-vector based on the average of i-vectors extracted from clean signals. A second approach relies on a vector quantizer codebook of representative clean speech i-vectors. Lastly, i-vectors and subjective ratings were used to train a no-reference deep neural network model for speech quality assessment. Four experiments have shown that the proposed methods, based on the i-vector speech representation, are well-suited for assessing speech quality. Results show correlations with subjective quality judgments similar to those achieved with standardized instrumental algorithms, particularly for degradations caused by noise and reverberation. \square

Keywords Speech quality assessment · Instrumental quality measurement · I-vector · Speech enhancement

Introduction

Estimating the perceived quality of existing and emerging multimedia services and applications is important, especially for providers seeking to optimize their services and maximize customer experience [1]. Real-time quality monitoring, for example, can help with network design and development, as well as with online adaptation to assure that the end users' expectations are met. As new services and technologies emerge, quality monitoring tools need to be able to characterize new artifacts and distortions that may arise.

Traditionally, subjective listening tests have been used and shown to be reliable [2]. In such a scenario, speech signals are presented to listeners (either naive or expert listeners, depending on the application) who judge the signal quality on a 5-point scale. The mean opinion score (MOS), which represents the perceived speech quality after leveling out individual factors [3], is attained after averaging all participant scores over a specific condition. Such subjective measurements, however, are not always feasible as they: (1) require many listeners; (2) can be laborious and time-consuming; (3) can be expensive; and (4) cannot be performed in real-time [4].

Instrumental quality measures, in turn, have been explored over the years to overcome these limitations. Instrumental measures are built to be highly correlated with subjective listening MOS scores, thus effectively replacing the listener panel by a real-time computational algorithm. These instrumental measures can be classified as full-reference, if they require a reference signal, or no-reference, if they operate only on the tested signal to compute the quality measurement. The International Telecommunication Union (ITU-T) has standardized several instrumental quality measures over the last couple of decades. The most widely recognized full-reference measure is the ITU-T Recommendation P.862, or

✉ Anderson R. Avila
anderson.avila@emt.inrs.ca

Jahangir Alam
jahangir@crim.ca

Douglas O'Shaughnessy
dougo@emt.inrs.ca

Tiago H. Falk
falk@emt.inrs.ca

¹ Institut national de la recherche scientifique, 800, rue de la Gauchetière Ouest, Montréal, QC H5A 1K6, Canada

² Computer Research Institut of Montreal, 405, Ogilvy Avenue, suite 101, Montréal, QC H3N 1M3, Canada

Perceptual Evaluation of Speech Quality (PESQ) [5], originally developed for narrow-band speech and later upgraded for wideband signals [6]. Recently, PESQ was superseded by ITU-T Recommendation P.863, or Perceptual Objective Listening Quality Prediction (POLQA) [7, 8], to account for emerging distortions, such as those seen with speech enhancement algorithms [9]. On the no-reference side, the latest method is the ITU-T Recommendation P.563 [10], which has been developed for narrow-band speech transmission networks. Recent research has shown that ITU-T P.563 does not perform well for conditions involving hands-free speech and speech enhancement algorithms [11, 12]; thus further innovations are still needed.

In this paper, we build upon the characteristics of the widely-used i-vector signal representation. The framework can be seen as a feature extraction procedure that depends basically on the observed speech signal, the universal background model (UBM) and the total variability matrix (or T matrix), which can be trained offline. We propose several new full-reference and no-reference instrumental quality measures applicable for hands-free speech, speech recorded in noisy real-world environments, and speech processed by enhancement algorithms. In particular, we expand the work in [13] to propose a full-reference measure based on the cosine similarity between the i-vector of a processed signal and the i-vector from its clean counterpart. Additionally, we explore three new no-reference variants, with two of them relying on different models of clean i-vector behaviour. I-vectors are computed from two different feature types, namely the traditional mel-frequency cepstral coefficients (MFCCs) and modulation spectral features (MSF). The latter have been shown useful in hands-free applications [14].

The main motivation behind this work lies in the fact that i-vectors are known to convey both channel and speaker information. Nevertheless, most research in the field has focused on the speaker characteristics of the representation (e.g., for speaker recognition) and channel effects have been usually discarded or overlooked. As shown in previous research [15, 16], however, the performance of i-vector based applications is severely affected by environmental factors, such as background noise and reverberation. To mitigate these channel effects, compensation techniques, such as linear discriminant analysis (LDA) and within class covariance normalization (WCCN) [17], are commonly applied. Here, unlike previous work, we utilize this information as a correlate of perceived speech quality. Moreover, as i-vectors are mapped to a fixed length feature vector, regardless of the originating signal length, full-reference quality assessment can bypass time-alignment, which is a crucial and error-prone step for PESQ and POLQA [3].

The remainder of this paper is organized as follows. [Background and proposed method](#) in section presents the proposed method and background on the i-vector framework.

[Experimental setup](#) in section describes the experimental setup and [“Experimental results and discussion”](#) in section presents the results and discussion. Lastly, conclusions are presented in [“Conclusions and future work”](#) section.

Background and proposed method

In this section, the proposed i-vector framework is presented. The steps to attain the features used to extract the i-vectors are also described. The measure used for estimating the distortion between reference and degraded signals is also discussed, as well as the effect of distortions on the i-vector representation.

The i-vector framework

The i-vector framework was developed inspired by the joint factor analysis (JFA) model [18]. The JFA can be seen as a Gaussian distribution of speaker and session-dependent supervectors. It is assumed that most of the variance in the supervector population is due to hidden variables, namely speaker and channel factors [19]. Both frameworks consider speaker and channel variability to lie in a low-rank subspace. Each component can be represented by a low-dimensional set of factors, which operate along the principal dimensions (also known as the eigenspace) of the corresponding component [20]. For JFA, this is represented as follows:

$$M = m + Vy + Ux + Dz \quad (1)$$

where m is the speaker- and channel-independent supervector, V the speaker eigen-voice matrix, D the diagonal residual matrix, U is the eigen-channel matrix, and y, z, x correspond to the low-dimensional eigen-voice, speaker-specific eigen-residual, and eigen-channel factors, respectively.

While the JFA approach models speaker and channel variability in separated subspaces, the i-vector framework considers only one subspace [15]. The argument for this new approach relies on the fact that channel factors estimated by JFA contain information about speakers, as shown in the experiments performed in [15]. The TV space, as for the JFA, is also defined by Gaussian mixture model (GMM) supervectors, which contain the mean values of a GMM universal background model (UBM) [20]. For instance, if the total number of mixture components is equal to K and the dimension of the acoustic feature vectors is denoted by F , then the supervector for a given recording is the concatenation of the mean vectors associated with the mixture components, which leads to a supervector of dimension KF . The supervector in the TV space can be represented as follows:

$$M = m + Tw, \quad (2)$$

where M is the speaker- and channel-dependent supervector extracted from a specific recording, m is the independent

supervector from the UBM, T corresponds to the total variability matrix trained with multiple recordings using the same procedure for learning the eigenvoice matrix [21], and w is a random vector with normal distribution, $N(0, I)$. This vector is referred to either as the identity vector or i-vector, and conveys the total factors [20]. Note that m can be seen as a prior probability distribution for the speaker supervectors, with the posterior probability distribution being derived from it in order to estimate speakers' dependent supervectors lying in the same eigenspace [21].

Figure 1 depicts the steps involving the extraction of i-vectors. Note that the framework ultimately maps a list of feature vectors into a fixed-length vector, $w \in \mathbb{R}^D$. These feature vectors, denoted here as $O = \{o_t\}_{t=1}^N$, where $o_t \in \mathbb{R}^F$, are extracted during the speech parameterization phase. In order to obtain w , a GMM model, $\lambda = (\{p_k\}, \{m_k\}, \{\sigma_k\})$, must be trained using multiple utterances. When such utterances come from different speakers, the model is referred to as a universal background model (UBM).

As depicted in Fig. 1, after training the GMM-UBM model, Baum-Welch statistics are extracted from each utterance u [22]. A total variability subspace is then learned and is used to estimate a low (and fixed) dimensional latent factor called the identity vector (i-vector) from adapted mean supervectors [23]. Note that the total factor w can be seen as the posterior distribution conditioned on the Baum-Welch statistics [15, 21], which is computed as follows:

$$N_k = \sum_{l=1}^L P(k|y_k, \lambda) \tag{3}$$

$$F_k = \sum_{l=1}^L P(k|y_k, \lambda)y_k \tag{4}$$

$$\tilde{F}_k = \sum_{l=1}^L P(k|y_k, \lambda)(y_k - m_k), \tag{5}$$

where the k -th frame is represented by y_k and L denotes the total number of frames extracted from a given utterance, and λ is the UBM. The mean of the k -th mixture component is represented by m_k . The posterior probability that the vector y_k is generated from the mixture component k is given by $P(k|y_k, \lambda)$. Note that Eq. (3) and Eq. (4) represent the zero-th and first-order Baum-Welch statistics, respectively. Eq. (5) is the centralized version of Eq. (4) [15]. The i-vector is then attained by

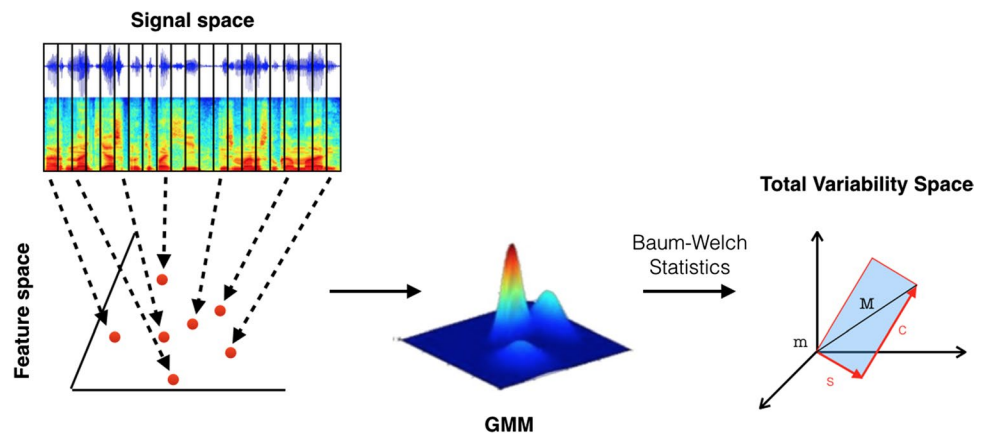
$$w = (I + T^T \Sigma^{-1} N(u) T)^{-1} T^T \Sigma^{-1} \tilde{F}(u), \tag{6}$$

with $N(u)$ being a diagonal matrix of $KF \times KF$ dimension and $\tilde{F}(u)$ is a supervector of dimension $KF \times 1$ obtained by the concatenation of first-order Baum-Welch statistics \tilde{F}_k for a given utterance u . A diagonal covariance matrix of $KF \times KF$ dimension is defined by Σ . Next, the two feature representations used to compute i-vectors are detailed.

Cepstrum parametrization

Mel-frequency cepstral coefficients (MFCC) have been widely used in speech applications for many years as they simulate the mel-scale present in the human cochlea [24]. Prior to their extraction, input speech signals re-sampled to 16 kHz are normalized to -26 dBOV. The signals also undergo a pre-emphasis filter of coefficient 0.95, which is meant to balance low and high frequency magnitudes. A 30-ms Hamming window with 50% overlap is applied before extracting the MFCCs. The Hamming window is used to remove edge effects [24]. The cepstral feature vector can then be extracted from each frame according to:

Fig. 1 Block diagram describing the steps for i-vector extraction



$$c_n = \sum_{m=1}^M [Y_m] \cos \left[\frac{\pi n}{M} \left(m - \frac{1}{2} \right) \right], n = 1, 2, 3, \dots, N, \tag{7}$$

where c_n is the n th mel-cepstral coefficient and Y_m refers to the log-energy of the m th filter. In this work, a set of 13 coefficients together with log energy, delta and delta-delta coefficients form the feature vector from each frame. As can be seen in (7), the MFCC representation is based on a short-term log-power spectrum and a cosine transformation on the nonlinear mel scale of frequency. Note that although cepstral normalization is commonly performed to minimize channel effects, it was not applied here, as we are interested in capturing such channel/distortion information.

Modulation spectrum parameterization

The modulation spectrum corresponds to an auditory spectro-temporal representation that captures long-term dynamics of the speech signal. It was shown in [26] that most speech content is concentrated under 20-Hz modulation frequency, whereas ambient artifacts, such as noise and reverberation, are beyond this threshold. Based on this insight, the authors proposed a measure called the speech-to-reverberation modulation energy ratio (SRMR), which was found to correlate well with reverberation levels and speech intelligibility. Motivated by their findings, we explore the use of modulation spectral features (MSF) for i-vector extraction.

To compute the modulation spectral features, we followed the same processing pipeline proposed by [26] and depicted in Fig. 2. During the pre-processing step, the speech activity level is normalized to -26 dBov (dB overload), thus eliminating unwanted energy variations caused by different loudness levels in the speech signal. Next, the pre-processed speech signal $\hat{x}(n)$ is filtered by a 23-channel gammatone filterbank, also simulating the cochlear processing [27]. The first filter of the filterbank is centered at 125 Hz and the last one at just below half of the sampling rate [26]. Each filter bandwidth follows the equivalent rectangular bandwidth (ERB) [27], which is an approximation of the bandwidths of the filters in human hearing, as described below:

$$ERB_j = \frac{f_j}{Q_{\text{ear}}} + B_{\text{min}}, \tag{8}$$

where f_j represents the center frequency of the j -th filter. Q_{ear} represents the asymptotic filter quality at high frequencies

and B_{min} is the minimum bandwidth for low frequencies. They are set, respectively, to 9.265 and 24.7.

The temporal envelope $e_j(n)$ is then computed from $\hat{x}_j(n)$, the output of the j -th acoustic filter, via the Hilbert transform:

$$e_j(n) = \sqrt{\hat{x}_j(n)^2 + \mathcal{H}\{\hat{x}_j(n)\}^2}, \tag{9}$$

where $\mathcal{H}\{\cdot\}$ denotes the Hilbert Transform. Temporal envelopes $e_j(n), j = 1, \dots, 23$ are then windowed with a 256-ms Hamming window and shifts of 40 ms. The discrete Fourier transform $\mathcal{F}\{\cdot\}$ of the temporal envelope $e_j(m;n)$ (m indexes the frame) is then computed in order to obtain the modulation spectrum $E_j(m, f_m)$, i.e.,

$$E_j(m; f_m) = |\mathcal{F}(e_j(m;n))|, \tag{10}$$

where m represents the m -th frame obtained after every Hamming window multiplication and f_m designates modulation frequency. The time variable n is dropped for convenience. Lastly, following recent physiological evidence of a modulation filterbank structure in the human auditory system [28], an auditory-inspired modulation filterbank is further used to group modulation frequencies into eight bands. These are denoted as $\mathcal{E}_{j,k}(m), k = 1, \dots, 8$, where j indexes the gammatone filter and k the modulation filter. Figure 3 depicts the frequency response for the 8-channel modulation filterbank used in our experiments. Note that the filter center frequencies are equally spaced in the logarithmic scale from 4 to 128 Hz.

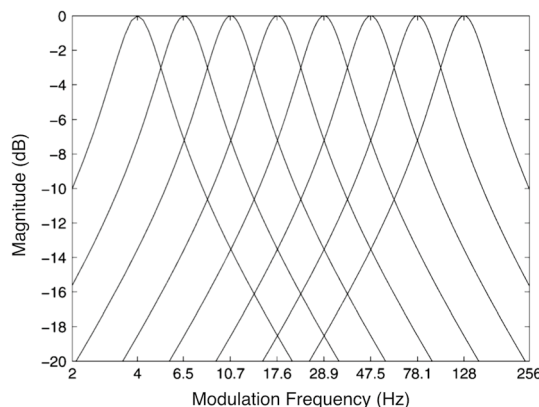


Fig. 3 Frequency responses of the 8-channel modulation filterbank, adapted from [25]



Fig. 2 Block diagram describing steps for computing the modulation spectrum representation

Cosine similarity scoring

The cosine similarity measure has been widely used to compare two supervectors in the total variability space [29]. It represents the angle between two total factor vectors, generated by (6) via the projection of two supervectors in the total variability space. The measure can be computed as follows:

$$\cos(\theta) = \frac{w_{ref} \cdot w_{deg}}{\|w_{ref}\| \cdot \|w_{deg}\|}, \tag{11}$$

where w_{ref} is the i-vector extracted from the reference speech recording and w_{deg} is the i-vector representation for the degraded speech recording. Figure 4 represents the total variability space, comprising the speaker and channel factors for two speech recordings of the same speaker. Considering that the proposed model is full-reference, we can assume no speaker and speech variability in the two representations. That is, speech content will remain the same for the reference and degraded signal and only changes in the channel factors will be present, as depicted in Fig. 4. Note that when significant alterations occur in the channel factors, the angle between w_{ref} and w_{deg} is expected to increase as well as the values of the cosine similarity. As such, the computation of the cosine similarity provides values close to 0 for high similarities and low distortions, and values close to 1 for low similarities and high distortions. Therefore, the similarities being captured are directly related to levels of distortions in the speech signal, as we show in the next section, and, thus, inversely proportional to speech quality.

Effects of distortions on the total variability subspace

In this section, we illustrate the impact of ambient noise on the i-vector representation to motivate our findings and hypotheses. For illustration purposes, we focus here only on reverberation and noise, as well as on MFCC features.

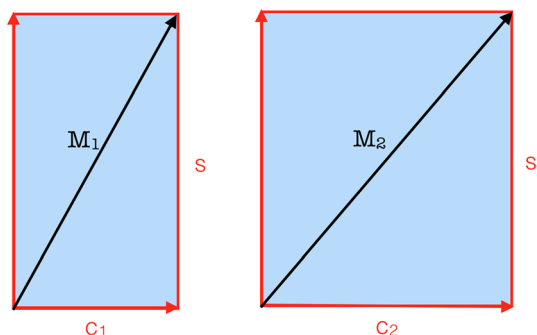


Fig. 4 Representation of speaker- and channel-dependent supervectors of two recordings from the same speaker where only the channel factors are affected

t-distributed stochastic neighbor embedding

In order to visualize the similarities between high-dimensional data points, it is convenient to map them into a two or three-dimensional space. Such projection must preserve the distances between data points, maintaining the structure of the high-dimensional data as much as possible. Note that in this process the interpretation of the coordinates becomes less important whereas the distances between data points and their clusterization carry out much more meaning. Here, we adopted a tool commonly used in machine learning, namely t-distributed stochastic neighbor embedding (t-SNE).

Different from reduction techniques, such as principal components analysis (PCA) that attempts to keep the low-dimensional representations of dissimilar data points far apart, the t-SNE method keeps the low-dimensional representations of very similar data points close together. According to [30], the t-SNE technique can capture local structure of the high-dimensional data and at the same time keep global structure such as the presence of clusters at several scales. To achieve this, the method embeds high-dimensional data into a lower-dimension space, usually two or three dimensions, by minimizing the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. The joint probability for the high-dimensional data can be expressed as:

$$p(x_j, x_i) = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq l} \exp(-\|x_k - x_l\|^2 / 2\sigma_i^2)}, \tag{12}$$

where x_i and x_j are data points and σ_i is the variance of the Gaussian distribution centered at x_i . Note that the conditional probability, $p(x_j|x_i)$, is assumed to be high for nearby data points, x_i and x_j , and low when x_i and x_j are far apart [30]. For the low-dimensional data, the joint probability takes the form of:

$$q(x_j, x_i) = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq l} \exp(-\|y_k - y_l\|^2)}, \tag{13}$$

where y_i and y_j are data points for data where we assume that the conditional probability, $q(x_j|x_i)$, is high for nearby data points, y_i and y_j , and low when y_i and y_j are far apart [30].

The cost function of the t-SNE is given by:

$$C = \sum_i KL(P_i || Q_i) = \sum_i \sum_j p(x_j, x_i) \log \frac{p(x_j, x_i)}{q(x_j, x_i)}, \tag{14}$$

where $p(x_i, x_i)$ and $p(y_i, y_i)$ are set to zero as modeling interests lie on pairwise similarities. It is important to mention that the t-SNE is an improved version of the Stochastic

Neighbor Embedding (SNE), which attempts to mitigate the so-called “crowded problem” encountered during the optimization [30].

Effects of background noise

Background noise plays an important role in the perception of speech quality. Therefore, it is expected that an instrumental quality measure will be sensitive to changes in signal-to-noise ratios (SNR’s). To summarize how the proposed model captures these changes, Fig. 5 depicts the effects of ambient noise on the TV subspace. For this, we added background noise at different levels (0, 5, 10, 15 dB) to clean speech files. In the figure, “Ref” stands for the corresponding reference clean speech signals. In order to visualize the similarities between data points (i.e., between i-vectors), we use t-SNE to embed high-dimensional i-vectors into two-dimensional space. Hence, each dot indicates an i-vector extracted from a speech recording and projected onto a two-dimensional space.

In Fig. 5, the recordings are labelled by SNR levels in the range of 0–15 dB (see different colors). The reference signal, which contains no distortions, is labelled as “Ref.” Note that the speech recordings with the same distortion levels are closely clustered. Moreover, as the SNR decreases, the clusters deviate from the clean speech cluster, with larger “distances” being seen for noisier cases. It is expected that the cosine similarity measure will be able to capture this distance information.

To give the reader more insights into the expected behaviour of the cosine similarity index, Fig. 7a provides the distribution of MOS as a function of SNR and Fig. 7b gives the cosine similarity as a function of SNR. As can be seen, cosine similarity is inversely proportional to SNR levels, which in turn are directly related to MOS.

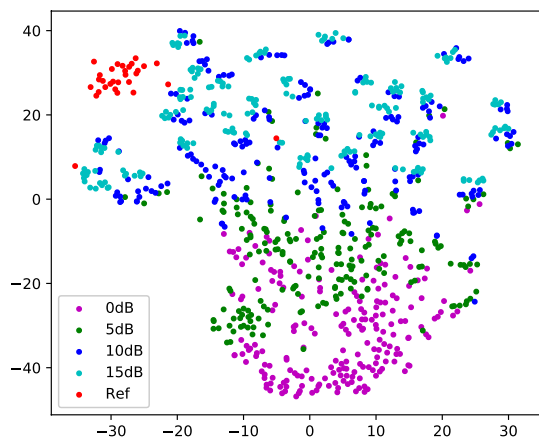


Fig. 5 i-Vector projection onto a 2-D space using t-SNE in the TV subspace at different levels of SNR

Effects of reverberation

Reverberation is characterized by the reflections of the speech signal on surfaces (e.g., walls) and objects present in an enclosed environment [31]. This directly changes the frequency response of the speech signal [32], which can have either positive or negative effects on the perceived quality of the speech. Early reflections, for instance, are desired as they cause changes in the signal timbre or coloration [32]. Late reflections, however, provide unwanted distortions represented by temporal smearing of the speech signal. As reverberation may affect the perceived quality of the speech [31], it is also expected that an instrumental quality measure will be able to rank different levels of reverberation, i.e., the time required for a signal to decay by 60 dB (also referred to as reverberation time or T_{60}) [33]. In Fig. 6, we show how different levels of reverberation are captured and represented in the TV space. For this, a small sample of clean speech files, each with different speech content, was convolved with a impulse response (IR) representing the following reverberation times: 0.25 s, 0.48 s and 0.8 s. In the figure, “Ref” stands for the corresponding reference clean speech signals. As previously, as reverberation levels increase, greater “distances” from the clean speech recordings are observed. Note, for instance, that the blue dots indicating recordings with $T_{60} = 0.80$ s are farther away from the red dots, which represent i-vector from clean recordings, than the purple dots, representing i-vectors from recordings with $T_{60} = 0.25$ s. This is expected as blue dots are the recordings with the highest amount of reverberation.

Label distribution versus SNR levels

To give the reader more insights of the expected behaviour of the proposed instrumental measure, Fig. 7 provides the

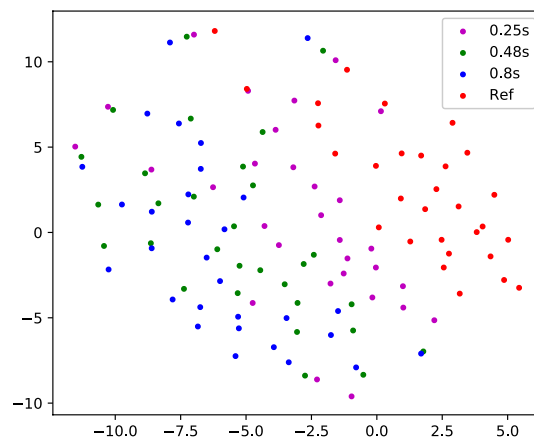


Fig. 6 i-vector disposition in the TV subspace at different levels of reverberation time (RT)

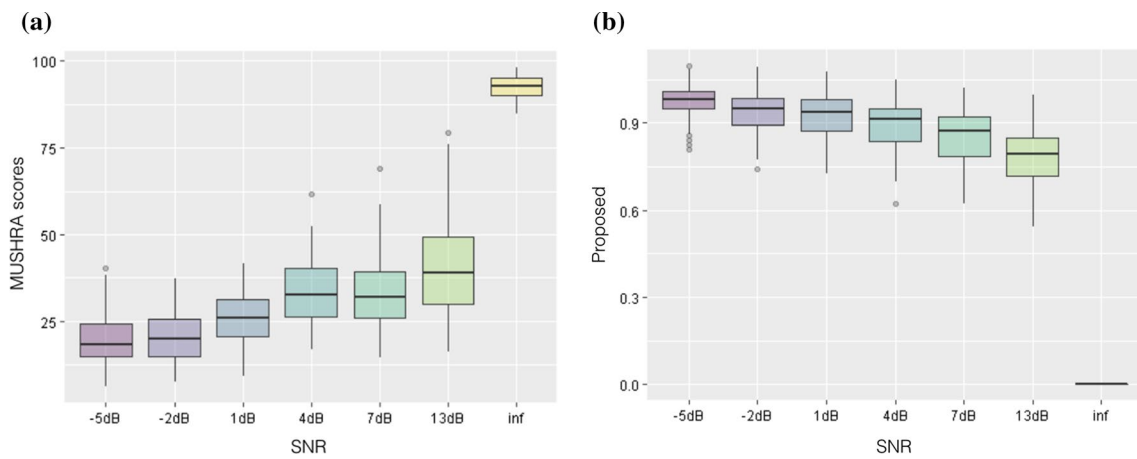


Fig. 7 Box-plot of (a) MUSHRA scores versus SNR and (b) cosine similarity versus SNR

distribution of MUSHRA scores [34] and cosine similarity according to values of SNR's, which range from -5 to 13 dB, including clean signal (see condition "inf"). For this, we used the INRS dataset, which will be described in "Data-base description" section. Note that, as mentioned before, cosine similarity is inversely proportional to perceived quality while the MUSHRA score is directly proportional to it. That is, while the MUSHRA score increases with the SNR, the proposed instrumental measure decreases. The best similarity is achieved with "inf", when no noise is present in the speech signal. This leads to the maximum MUSHRA scores of 100 and minimum cosine similarity of 0 , as can be seen in Fig. 7b. Moreover, we observe a close trend in both distributions. For example, for the two lowest SNR's (i.e. -5 dB and -2 dB) we also have the two lowest MOS, followed by a slight improvement on the perceived quality. We see a similar pattern with the proposed method, but with maximum values of cosine similarity for low SNR's. Of

course, we cannot assume by this that MUSHRA scores and our method are then correlated. However, we show in Fig. 8 that there is a clear trend between our predictions and the MUSHRA scores for the same database. Note that Fig. 8a is the scatterplot for different levels of SNR while Fig. 8b is the scatterplot for different reverberation times.

Proposed methods

Given the insights mentioned above, four new instrumental measures are proposed, one full-reference and three no-reference, as detailed next.

Full-reference

This approach relies on the cosine similarity between i-vectors extracted from the reference (clean) signal and i-vectors

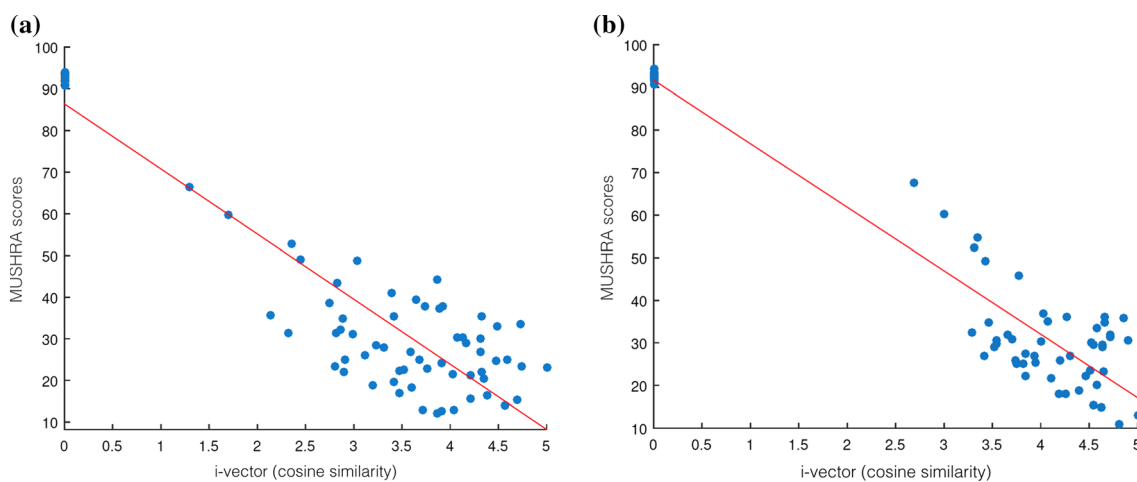


Fig. 8 Scatter-plots of MUSHRA scores versus cosine similarity metric for speech corrupted with (a) only noise and (b) only reverberation

extracted from their degraded speech signal counterparts. This similarity index is used as a correlate of speech quality.

No-reference: average reference model

As no-reference models do not have access to a reference signal, models of clean speech are required. The first approach proposed here models clean speech as an average i-vector computed from clean speech. More specifically, i-vectors are extracted from clean speech data (see [Database description](#) in section) and averaged to obtain one reference i-vector always to be used in the computation of the cosine similarity metric. As discussed in [Effects of distortions on the total variability subspace](#) section, larger distances from this average i-vector should indicate lower quality signals.

No-reference: vector quantizer codebook reference model

Inspired by earlier works on instrumental speech quality measurement [35], the second proposed approach relies on a vector quantizer (VQ) codebook of reference i-vectors obtained from clean speech. This builds upon the previous “average i-vector” method in that a different reference i-vector is used for each processed signal. In particular, the i-vector that most closely resembles the degraded signal i-vector is used for computation of the cosine similarity distance. Here, the k -means algorithm is used to build the vector quantization codebook. We tested different values of k in the range of 5–500, and found that the optimal number usually represented the number of distortions in the datasets used in our experiments.

No-reference: deep neural network reference model

Lastly, we are inspired by recent innovations in no-reference methods based on deep neural networks [12, 36]. Here, a DNN is trained to estimate MOS. In particular, a fully-connected model with 400 input units (i.e., it receives i-vectors with 400 factors) and three hidden layers is used. The first hidden layer has 200 units, followed by 100 and 50 units. We adopted ReLU as the activation function and the output unit is a simple linear function. We used dropout with 0.2 rate and Adadelta as the optimization method. Dropout function is used as a regularizer to avoid over-fitting, and the Adadelta optimizer is used to dynamically update the learning rate and is suited to sparse data [37].

Experimental setup

In this section, the databases used in our experiments are presented, followed by a discussion on comparing subjective and objective speech quality scores. Then, details about the

i-vector extraction are given and a description of the figures-of-merit used is presented.

Database description

Three main datasets are used herein: (1) the noise speech database developed by [38], (2) the INRS audio quality dataset [39], and (3) the open-source noise speech corpus NOIZEUS [40]. The noise speech database is a clean and noisy parallel speech dataset, developed for the purpose of training speech enhancement algorithms, such as the speech enhancement generative adversarial network (SEGAN) [41]. It contains pairs of clean and noisy speech samples from 28 speakers (14 males and 14 females), all from the same accent region (England), taken from the larger Voice Bank corpus [42]. The dataset is sampled at 48 kHz. For the purpose of our experiments, only clean utterances were used from this dataset and solely to train the i-vector framework (i.e., the GMM-UBM and total variability matrix), as well as two of our proposed no-reference approaches: (1) the one based on the reference i-vector, attained from the average of the extracted clean signals, and (2) the one based on the reference VQ codebook, also attained from the extracted clean signals. It is important to mention that the next two datasets are adopted to assess the models and none of their samples are used for training, except for the last experiment where a no-reference neural network based model is evaluated.

The INRS dataset, in turn, contains speech files sampled at 16 kHz and degraded by noise and reverberation. To this end, clean signals from the TIMIT database were corrupted with babble and factory noises at SNRs of -2 dB, -5 dB, 1 dB, 4 dB, 7 dB and 13 dB. Noise signals were obtained from the NOISEX-92 dataset [43]. Reverberant utterances, in turn, were generated by convolving the clean utterances with 740 room impulse responses (RIR) with the following reverberation times (T_{60}): 0.3 s, 0.6 s, 0.9 s, 1.2 s and 1.5 s. For each T_{60} value, twenty different simulated RIRs (with different room geometry, source microphone positioning and absorption characteristics) were used. The RIRs were generated using an image-source method tool for simulating sound fields in virtual reverberant environments [44]. From the noisy signals, three speech DNN-based enhancement models were used. The first uses a feed-forward neural network to estimate the spectral data. The second model proposes the use of a feed-forward neural network in combination with arbitrary features to estimate a spectral mask. The final investigated enhancement model is based on spectral estimation through a context-aware recurrent neural network model. Details about these algorithms can be found in [39].

The listening test followed the MUSHRA methodology. Two different online listening tests are performed, one for the dereverberation (112 participants and 10 conditions) and one for the noise suppression (245 participants and 12

conditions). Both tests were MUSHRA-style tests where the output of all models (i.e., dereverberation or noise suppression), a hidden reference, a corrupted anchor, and the corrupted signals were presented to each participant. A slider with their positions quantized as integers ranging from 0 to 100 was used by the listeners to rate the signal quality. For each noise type, an anchor was the same stimulus corrupted with a 5 dB lower SNR. For the dereverberation conditions, in turn, the anchor was a signal convolved with an RIR with a T_{60} of 2 s. More details can be found in [39].

Lastly, the NOIZEUS database contains 30 IEEE sentences recorded by three male and three female speakers in a sound-proof booth. The speech signals, sampled at 8 kHz, are contaminated with eight different noise types taken from the AURORA database [45] and include car, train, babble, exhibition, restaurant, street, airport and station noises. Noise is added to the clean speech signals at SNRs of 0, 5, 10 and 15 dB. In our experiments, a subset of NOIZEUS is considered and includes 4 noise types (i.e., babble, car, street and train) and two SNR levels (i.e., 5 and 10 dB). Thirteen speech noise-suppression algorithms are also applied to the corrupted samples; a complete list is available in [45].

The subjective test conducted was based on the ITU-T Recommendation P.835, which aims at reducing the listeners' uncertainty to which component (i.e., the speech signal, the background noise, or both) to take into account when rating the signal quality. To this end, listeners are instructed to first evaluate the speech signal alone using a five-point scale of signal distortion (SIG). Next, they attend to the background noise alone using a five-point background intrusiveness (BAK) scale, and lastly, they are instructed to focus on the overall effect using the five-point mean opinion score (OVRL) – [1 = bad, 2 = poor, 3 = fair, 4 = good, 5 = excellent]. A total of 32 participants, between the ages of 18 and 50, took part in the test. The interested reader is referred to [45] for more details about the subjective listening test, including the details regarding the SIG and BAK scales.

Comparing subjective and objective speech quality scores

Subjective listening tests can follow different scales. For example, ITU-T P.800 [2] describes the five-point absolute category rating (ACR) scale with ratings varying from bad to excellent, respectively. MUSHRA tests, on the other hand, rely on a 0–100 scale with lower values corresponding to poor quality and higher values to excellent. While the different scales usually rank similarly, direct comparisons between them needs to be done carefully. Moreover, certain instrumental algorithms may be tuned and calibrated to a specific rating scale, thus comparing their outputs to any arbitrary subjective scale may be misleading. To avoid such issues, here we propose to use a third order monotonic polynomial

mapping [46] to map objective and subjective ratings into the same scale prior to computation of the figures-of-merit. This mapping is applied to both proposed and benchmark instrumental measures.

i-Vector framework setup

Once speech parameterization was performed, a Gaussian mixture model universal background model (GMM-UBM) with 1024 Gaussians was trained using 824 clean utterances from the noise speech database [42]. Five different total variability matrix T sizes were explored, namely containing 100, 200, 300, 400 or 500 total factors. The motivation behind testing different TV subspaces was to obtain insight on what number of factors is optimal for speech quality assessment. After training, i-vectors were extracted from all recordings from the INRS and NOIZEUS datasets.

Figures-of-merit and benchmark algorithms

Here, the performance of the proposed and benchmark instrumental measures are assessed using two figures-of-merit. The linear relationship between estimated quality and ground truth is computed using the Pearson correlation coefficient (ρ_{Person}). We also consider the root mean-square error (RMSE) [47] to represent the error between predicted values and ground truths. As mentioned previously, figures-of-merit are computed after a third-order monotonic mapping is applied to guarantee that instrumental and subjective ratings are in the same scale.

Correlations are reported on a per-condition basis, where all files under the same acoustic condition are first averaged prior to correlation calculation. Table 1 describes the acoustic conditions present in the INRS dataset excluding references and anchors. Two noise types (babble and factory) are considered at six different SNR levels. Moreover, three enhancement algorithms were applied, referred to herein as Santos2018 [48], Williamson2017 [49] and Wu2016 [50]. Thus, for the noisy samples, an acoustic condition is defined by the noise type, processing status and SNR level, while for the reverberant samples, an acoustic condition is configured by the processing status and the reverberation time (T_{60}). The NOIZEUS dataset, on the other hand, presents four noise types: babble, car, street and train at two different SNR levels. For noise suppression, two algorithms based on the Wiener filter are tested, referred herein as Wavthre [51] and Tsoukalas [52]. Table 2 summarizes the conditions in the NOIZEUS database.

Lastly, to gauge the advantages of the proposed methods, two standard full-reference methods are used as benchmarks, i.e., wide-band mode ITU-T P.862.2 (PESQ) [6] and full-band mode ITU-T P.863 (POLQA) [7]. For the no-reference measures, the narrow-band ITU-T P.563

Table 1 Overview of the INRS speech quality dataset where acoustic conditions are presented for denoising and dereverberation processes, excluding reference files and anchors

Denoising		Dereverberation		
Noise type	Enhancement algorithms	SNR (dB)	Enhancement algorithms	T_{60}
Babble, Factory	Unprocessed, Santos2018, Williamson2017, Wu2016	- 5	Unprocessed, Santos2018, Williamson2017, Wu2016	0.3 s
		- 2		0.6 s
		1		0.9 s
		4		1.2 s
		7		1.5 s
		13		

Table 2 Overview of the NOIZEUS speech quality dataset where acoustic conditions are presented for denoising, excluding references

Noise type	Enhancement algorithms	SNR
Babble, Car, Street, Train	Unprocessed, Wavthre, Tsoukalas	5 dB 10 dB

[10] and wide-band SRMR [31] are used as benchmarks. It is important to emphasize that no standardized wide-band no-reference measure exists; thus the results reported herein for ITU-T P.563 are at somewhat of a disadvantage, as the proposed method considers wide-band signals. For POLQA, a similar consideration must be made as the model has been used with wide-band signals, which is out of normal operation for the model. Notwithstanding, comparisons with PESQ and SRMR provide a more fair performance comparison to gauge the benefits of the proposed method.

Experimental results and discussion

In this section, we describe our experiments and provide a discussion of the achieved results.

Experiment I: Full-reference measurement

Table 3 presents the per-condition performances on the INRS database for the noise-only and reverberation-only settings (each condition also includes the enhanced counterpart), as well as the performances on the NOIZEUS dataset for noisy and enhanced speech. In the table, the performances of the two benchmark full-reference algorithms are also presented for comparison purposes. As can be seen, i-vectors extracted from MSFs are able to better correlate with subjective ratings for the noise, reverberation, and enhanced conditions, whereas MFCC-based ones are not as effective for the enhancement case. For noise and reverberation conditions, a larger number of factors (400-500) resulted in the best results, whereas for enhancement alone,

Table 3 Per-condition performance of the proposed full-reference approach on the INRS and NOIZEUS databases. Numbers in subscript indicate the number of factors in the total variability space

Metrics	INRS (MUSHRA)				NOIZEUS (ACR)			
	Noise		Reverb		Noise		Enhanced	
	$\rho_{Pearson}$	$RMSE$	$\rho_{Pearson}$	$RMSE$	$\rho_{Pearson}$	$RMSE$	$\rho_{Pearson}$	$RMSE$
PESQ	0.95	7.32	0.92	9.58	0.93	0.12	0.90	0.13
POLQA	0.96	6.70	0.92	9.73	0.89	0.16	0.92	0.14
MFCC ₁₀₀	0.94	8.42	0.90	11.16	0.98	0.20	0.49	0.28
MFCC ₂₀₀	0.94	8.22	0.92	9.58	0.94	0.21	0.50	0.28
MFCC ₃₀₀	0.94	8.17	0.92	9.90	0.97	0.21	0.60	0.28
MFCC ₄₀₀	0.94	8.12	0.92	9.59	0.81	0.21	0.04	0.30
MFCC ₅₀₀	0.95	7.97	0.92	9.55	0.76	0.28	0.02	0.30
MSF ₁₀₀	0.93	8.96	0.94	8.06	0.97	0.11	0.82	0.22
MSF ₂₀₀	0.89	11.78	0.94	8.24	0.98	0.12	0.82	0.24
MSF ₃₀₀	0.93	9.43	0.93	8.68	0.98	0.13	0.87	0.22
MSF ₄₀₀	0.94	8.10	0.93	8.80	0.71	0.31	0.24	0.29
MSF ₅₀₀	0.94	8.37	0.93	8.84	0.83	0.28	0.61	0.27

Bold values indicate the best performance

a smaller number of factors sufficed (200). Overall, MFCC-based i-vectors showed to be more sensitive to the number of factors relative to MSF-based ones. Correlation values were in line with those obtained with the benchmarks, but with the added benefit of not requiring temporal alignment. In the case of the NOIZEUS dataset, the proposed method achieved slightly higher correlations, i.e., 0.98 for the noisy condition, when compared to the benchmarks. However, the performance of the proposed method (0.87) was not as good for the enhancement condition, with the best correlation, 0.92, being achieved by POLQA.

Experiment II: No-reference measurement based on average model

Table 4 shows the results attained using the simplest no-reference measure, as well as the two no-reference benchmarks. As expected, results are much lower than what can be achieved with a full-reference measure. Notwithstanding, the simple average i-vector model extracted from MSFs outperformed both benchmarks on the INRS database when 100, 200 and 300 factors were adopted. The simple approach, however, was not capable of accurately tracking the quality of the enhanced signals in the NOIZEUS database, despite outperforming ITU-T P.563. I-vectors extracted from MFCCs, however, were able to quantify the distortions in the noise-only case for the NOIZEUS database and results in line with SRMR were achieved.

Overall, reverberation showed to be a harder problem. This was expected based on insights from Figs. 5 and 6. In fact, the projection of i-vectors is less clustered in the case of reverberation, thus making it harder to model using such a simple approach. Furthermore, it is interesting to note that,

in this case (MSF features), performance is inversely proportional to the number of factors.

Experiment III: No-reference measurement based on VQ codebook

Table 5 provides results obtained with the VQ codebook-based approach. We tested different numbers of clusters (i.e., 10, 30 and 40) and we adopted 10 clusters, which seemed to be the optimal value. The performances are presented for both MFCCs and MSF-based systems. We can note that the proposed metric based on MSF-k10 provides more stable results throughout the tested conditions. Moreover, it presents competitive performance compared to the two benchmarks, SRMR and P.563, outperforming the latter for all tested conditions and SRMR in two situations. See the first and last columns with the respective correlations equal to -0.72 and -0.80 . We can verify that distortions caused by reverberation and enhanced speech are more challenging to the proposed metric. In fact, the results were more reliable for distortions caused by noise for all the metrics, including the proposed MSF-k10.

Experiment IV: No-reference measurement based on DNNs

In this experiment, the model is trained considering both noisy and reverberant samples. We randomly sampled 70% of the examples in the INRS database to train. During training, 20% of these examples are used for validation. The remainder of the dataset is kept for testing. The results presented in Table 6 are based on the average value after running the experiment 10 times, randomly picking samples

Table 4 Per-condition performance of the no-reference approach based on average of i-vectors for the INRS and NOIZEUS databases. Numbers in subscript indicate the number of factors in the Total Variability Space

Metrics	INRS (MUSHRA)				NOIZEUS (ACR)			
	Noise		Reverb		Noise		Enhanced	
	$\rho_{Pearson}$	RMSE	$\rho_{Pearson}$	RMSE	$\rho_{Pearson}$	RMSE	$\rho_{Pearson}$	RMSE
SRMR	0.70	20.55	0.48	23.01	0.89	0.24	0.72	0.32
P563	0.48	23.47	0.38	25.37	0.87	0.25	0.13	0.30
MFCC ₁₀₀	0.13	25.63	0.19	24.94	0.81	0.28	0.04	0.30
MFCC ₂₀₀	0.28	25.41	0.09	25.34	0.76	0.27	0.02	0.30
MFCC ₃₀₀	0.63	20.78	0.22	24.76	0.39	0.30	0.03	0.30
MFCC ₄₀₀	0.06	25.75	0.17	25.15	0.46	0.30	0.17	0.30
MFCC ₅₀₀	0.57	23.39	0.20	24.94	0.43	0.30	0.77	0.29
MSF ₁₀₀	0.90	12.97	0.67	19.78	0.82	0.28	0.24	0.29
MSF ₂₀₀	0.84	16.84	0.38	23.57	0.83	0.28	0.61	0.27
MSF ₃₀₀	0.80	17.54	0.61	20.95	0.37	0.31	0.53	0.27
MSF ₄₀₀	0.55	22.84	0.44	23.20	0.19	0.31	0.58	0.28
MSF ₅₀₀	0.53	23.89	0.44	23.50	0.58	0.29	0.17	0.29

Bold values indicate the best performance

Table 5 Per-condition performance of the no-reference approach based on a VQ codebook for the INRS and NOIZEUS databases

Metrics	INRS (MUSHRA)				NOIZEUS (ACR)			
	Noise		Reverb		Noise		Enhanced	
	$\rho_{Pearson}$	$RMSE$	$\rho_{Pearson}$	$RMSE$	$\rho_{Pearson}$	$RMSE$	$\rho_{Pearson}$	$RMSE$
SRMR	0.70	20.55	0.48	23.01	0.89	0.24	0.72	0.32
P563	0.48	23.47	0.38	25.37	0.87	0.25	0.13	0.30
MFCC _{k10}	0.09	25.74	0.31	24.54	0.95	0.21	0.61	0.29
MFCC _{k30}	0.06	25.75	0.48	24.77	0.93	0.21	0.59	0.29
MFCC _{k40}	0.57	23.39	0.14	25.75	0.87	0.24	0.51	0.30
MSF _{k10}	0.78	21.26	0.47	23.31	0.94	0.29	0.50	0.30
MSF _{k30}	0.55	22.84	0.52	24.81	0.07	0.30	0.55	0.29
MSF _{k40}	0.53	23.89	0.60	24.16	0.92	0.27	0.29	0.29

Bold values indicate the best performance

Table 6 Per-condition performance of a no-reference DNN-based model trained with i-vector features from speech samples from the INRS database

Metrics	$\rho_{Pearson}$	$RMSE$
PESQ	0.90	10.06
POLQA	0.89	10.05
MFCC ₁₀₀	0.77	15.11
MFCC ₂₀₀	0.82	13.17
MFCC ₃₀₀	0.90	10.43
MFCC ₄₀₀	0.94	7.60
MFCC ₅₀₀	0.90	10.25
MSF ₁₀₀	0.85	12.29
MSF ₂₀₀	0.86	12.17
MSF ₃₀₀	0.88	11.29
MSF ₄₀₀	0.87	11.44
MSF ₅₀₀	0.85	12.49

Bold values indicate the best performance

for the training and test sets. As can be seen, the best results are achieved by the MFCC-based i-vector with 400 factors, where correlations are higher than those achieved with PESQ and POLQA. Models based on MSFs and 300 factors achieved similar results to PESQ and POLQA. It is important to emphasize that the obtained results may be somewhat optimistic, as the proposed method relied on a subset of the INRS database for training, whereas the benchmarks did not. Nonetheless, the achieved results are promising, as they show a no-reference method achieving comparable results to full-reference benchmarks.

Conclusions and future work

In this paper, we explored the use of i-vector speech representations for instrumental quality measurement of noisy, reverberant and enhanced speech. The UBM and T matrix are normally trained offline. We show how the total

variability space is capable of capturing ambient factors and one full-reference and three no-reference measures are proposed. Experimental results on two datasets showed the full-reference method achieving results in line with two standard benchmarks and bypassing the need for time alignment between reference and processed signals. On the same datasets, the three no-reference measures presented higher correlations with subjective quality scores compared to two no-reference benchmarks, thus showing their effectiveness in tracking the quality of hands-free and enhanced speech. As future work, we intend to investigate other strategies to improve the performance of the i-vector framework as a no-reference speech quality metric. We also intend to test the proposed method on an extended dataset with network impairments; of particular interest is the behaviour of the proposed method under temporal distortions.

Acknowledgements The authors would like to thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), the Fonds de recherche du Québec - Nature et Technologies (FRQNT), and the Natural Sciences and Engineering Research Council of Canada (NSERC) for their financial support.

References

1. Wu D et al (2015) Millimeter-wave multimedia communications: challenges, methodology, and applications. *IEEE Commun Mag* 53(1):232–238
2. ITU-T. Recommendation P.800 (1998) Methods for subjective determination of transmission quality
3. Moller S et al (2006) Speech quality estimation: models and trends. *IEEE Signal Process Mag* 28(6):18–28
4. Avila A R et al (2016) Performance comparison of intrusive and non-intrusive instrumental quality measures for enhanced speech. IWAENC
5. ITU-T. Recommendation P.862 (2001) Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs

6. ITU-T (2007) Recommendation P.862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs
7. Recommendation P.863
8. ITU-T. Recommendation P.863 (2018) Perceptual objective listening quality prediction: telephone transmission quality, telephone installation, local line networks—methods for objective and subjective assessment of speech quality
9. Falk TH et al (2015) Objective quality and intelligibility prediction for users of assistive listening devices: advantages and limitations of existing tools. *IEEE Signal Process Mag* 32(2):114–124
10. Malfait L, Berger J, Kastner M (2006) P. 563 the ITU-T standard for single-ended speech quality assessment. *IEEE Trans Audio Speech Lang Process* 14(6):1924–1934
11. Avila A R et al (2016) Performance comparison of intrusive and non-intrusive instrumental quality measures for enhanced speech. In: 2016 IEEE international workshop on acoustic signal enhancement (IWAENC), pp 1–5. IEEE
12. Avila A R et al (2019) Non-intrusive speech quality assessment using neural networks. In: International conference on acoustics, speech and signal processing (ICASSP), pp. 631–635. IEEE,
13. Avila AR et al (2019) Intrusive quality measurement of noisy and enhanced speech based on i-vector similarity. In: 2019 Eleventh international conference on quality of multimedia experience (QoMEX), pp 1–5. IEEE
14. Falk T, Chan WY (2009) Modulation spectral features for robust far-field speaker identification. *IEEE Trans Audio Speech Lang Process* 18(1):90–100
15. Dehak N et al (2011) Front-end factor analysis for speaker verification. *IEEE Trans Audio Speech Lang Process* 19(4):788–798
16. Garcia-Romero D, Zhou X, Espy-Wilson CY (2012) Multicondition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4257–4260. IEEE
17. Dehak N et al (2010) Cosine similarity scoring without score normalization techniques. In: *Odyssey*, pp. 15
18. Kenny P et al (2007) Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Trans Audio Speech Lang Process* 15(4):1435–1447
19. Kenny P (2005) Joint factor analysis of speaker and session variability: theory and algorithms. In: CRIM, Montreal,(Report) CRIM-06/08-13, vol 14, pp 28–29
20. Hansen JHL, Hasan T (2015) Speaker recognition by machines and humans: a tutorial review. *IEEE Signal Process Mag* 32(6):74–99
21. Kenny P, Boulianne G, Dumouchel P (2005) Eigenvoice modeling with sparse training data. *IEEE Trans Speech Audio Process* 13(3):345–354
22. Garcia-Romero D, Espy-Wilson CY (2011) Analysis of i-vector length normalization in speaker recognition systems. In: Twelfth annual conference of the international speech communication association
23. Sadjadi SO, Slaney M, Heck L (2013) Msr identity toolbox v1. 0: a matlab toolbox for speaker-recognition research. *Speech Lang Process Tech Comm Newslett* 1(4):1–32
24. Logan B et al (2000) Mel frequency cepstral coefficients for music modeling. *Ismir* 270:1–11
25. Falk TH, Chan WY (2010) Temporal dynamics for blind measurement of room acoustical parameters. *IEEE Trans Instrum Meas* 59(4):978–989
26. Falk TH, Chan WY (2010) Modulation spectral features for robust far-field speaker identification. *IEEE Trans Audio Speech Lang Process* 18(1):90–100
27. Slaney M et al (1993) An efficient implementation of the patten-son-holdsworth auditory filter bank. Apple Computer, Perception Group. Tech Rep 35:8
28. Ewert SD, Dau T (2000) Characterizing frequency selectivity for envelope fluctuations. *J Acoust Soc Am* 108(3):1181–1196
29. Shum S et al (2010) Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification. In: *Odyssey*, pp 16
30. Laurens LVM, Hinton G (2008) Visualizing data using t-sne. *J Mach Learn Res* 9:569
31. Falk TH, Zheng C, Chan WY (2010) A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Trans Audio Speech Lang Process* 18(7):1766–1774
32. Halmrast T (2001) Sound coloration from (very) early reflections. *J Acoust Soc Am* 109(5):2303
33. Joyce WB (1975) Sabine's reverberation time and ergodic auditoriums. *J Acoust Soc Am* 58(3):643–655
34. ITU-R Rec. Itu-r bs. 1534-1 (2003) Method for the subjective assessment of intermediate quality level of coding systems
35. Jin C, Kubichek R (1996) Vector quantization techniques for output-based objective speech quality. In: 1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings, vol 1, pp 491–494. IEEE
36. Cauchi B et al (2019) Non-intrusive speech quality prediction using modulation energies and lstm network. *IEEE/ACM Trans Audio Speech Lang Process* 27(7):1151–1163
37. Ruder S (2016) An overview of gradient descent optimization algorithms. arXiv preprint [arXiv:1609.04747](https://arxiv.org/abs/1609.04747)
38. Valentini-Botinhao C et al (2017) Noisy Speech Database for Training Speech Enhancement Algorithms and tts Models. University of Edinburgh. School of Informatics, Centre for Speech Technology Research (CSTR), Edinburgh
39. Santos J, Falk TH (2019) Towards the development of a non-intrusive objective quality measure for dnn-enhanced speech. In: 2019 eleventh international conference on quality of multimedia experience (QoMEX), pp. 1–6. IEEE
40. Hu Y, Loizou PC (2007) Subjective comparison and evaluation of speech enhancement algorithms. *Speech Commun* 49(7–8):588–601
41. Pascual S, Bonafonte A, Serrà J (2017) Segan: Speech enhancement generative adversarial network. arXiv preprint [arXiv :1703.09452](https://arxiv.org/abs/1703.09452)
42. Veaux C, Yamagishi J, King S (2013) The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In: 2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE), pp. 1–4. IEEE
43. Varga A, Steeneken HJM (1993) Aassessment for automatic speech recognition: Ii. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Commun* 12(3):247–251
44. Lehmann EA, Johansson AM (2009) Diffuse reverberation model for efficient image-source simulation of room impulse responses. *IEEE Trans Audio Speech Lang Process* 18(6):1429–1439
45. Hirsch H, Pearce D (2000) The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: ASR2000—automatic speech recognition: challenges for the new Millennium ISCA tutorial and research workshop (ITRW)
46. Rix AW (2003) Comparison between subjective listening quality and p. 862 pesq score. In: Proceedings measurement of speech and audio quality in networks (MESAQIN03), Prague, Czech Republic
47. Shcherbakov MV et al (2013) A survey of forecast error measures. *World Appl Sci J* 24(24):171–176

48. Santos JF, Falk TH (2018) Speech dereverberation with context-aware recurrent neural networks. *IEEE/ACM Trans Audio Speech Lang Process* 26(7):1236–1246
49. Williamson DS, Wang D (2017) Time-frequency masking in the complex domain for speech dereverberation and denoising. *IEEE/ACM Trans Audio Speech Lang Process* 25(7):1492–1501
50. Wu B et al (2016) A reverberation-time-aware approach to speech dereverberation based on deep neural networks. *IEEE/ACM Trans Audio Speech Lang Process* 25(1):102–111
51. Hu Y, Loizou PC (2004) Speech enhancement based on wavelet thresholding the multitaper spectrum. *IEEE Trans Speech Audio Process* 12(1):59–67
52. Tsoukalas DE, Mourjopoulos JN, Kokkinakis G (1997) Speech enhancement based on audible noise suppression. *IEEE Trans Speech Audio Process* 5(6):497–514

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.