Automatic pathological speech intelligibility assessment exploiting subspace-based analyses

Parvaneh Janbakhshi, *Student Member, IEEE,* Ina Kodrasi, *Member, IEEE,* and Hervé Bourlard, *Fellow, IEEE*

***Abstract*—Competitive state-of-the-art automatic pathological speech intelligibility measures typically rely on regression training on a large number of features, require a large amount of healthy speech training data, or are applicable only to phonetically balanced scenarios where healthy and pathological speakers utter the same utterances. As a result, their performance in unseen data is unsatisfactory, and they cannot be used in low-resource languages or in phonetically unbalanced scenarios. To overcome these drawbacks, we propose a subspace-based intelligibility (SBI) measure. The SBI measure operates based on the hypothesis that dominant spectral patterns of pathological speech differ from intelligible speech (where the pathological and intelligible speech signals do not need to match in phonetic content), with the difference increasing as pathological speech intelligibility decreases. The SBI measure uses a minimal number of speech recordings to compute dominant spectral basis vectors spanning intelligible and pathological speech. The subspaces spanned by the intelligible and pathological spectral basis vectors are compared to each other through a subspace distance measure, which is directly used (i.e., without any training) as the pathological speech intelligibility estimate. Exploiting psychoacoustic evidence on the importance of spectral modulation cues to the perceived speech intelligibility and clinical evidence on the degradation of these cues in pathological speech, we show that the power of the proposed SBI measure lies in capturing the effect of spectral modulation degradation. To be able to additionally track possible degradations in the temporal structure of the pathological speech signal, we also propose two extensions of the SBI measure by incorporating short-time temporal information. Experimental results for different languages and speech pathologies show that the proposed intelligibility measures yield high and significant correlations with subjective intelligibility ratings, while not requiring any regression training or a large number of healthy speech recordings and being applicable to phonetically unbalanced scenarios.**

***Index Terms*—Spectral subspace, SVD, spectral modulation, Cerebral Palsy, hearing impairment**

I. Introduction

SPEECH intelligibility assessment is an important component of the auditory-perceptual evaluation of pathological speech in clinical settings, since it can be used to characterize the severity of the speech disorder and to monitor the effectiveness of speech therapy. Clinical approaches to pathological speech intelligibility assessment are based on subjective listening tests where human listeners directly evaluate the perceived speech intelligibility [1]. Such subjective tests are time-consuming and may be biased by the availability of syntactic/semantic clues in connected speech as well as by the familiarity of the listener with the speaker, speech disorder, or speech task under study [2], [3]. As an efficient and economical substitute to subjective intelligibility assessment, automatic pathological speech intelligibility measures have been proposed. Such objective measures avoid the above-mentioned drawbacks of subjective listening tests and offer repeatable assessments with the capability of being used in real-time and remote speech therapy applications [4].

Approaches to automatic pathological speech intelligibility assessment can be broadly categorized into blind approaches [5]–[9] and non-blind approaches [10]–[20]. Blind approaches refer to approaches that do not exploit any knowledge about healthy (i.e., intelligible) speech and assess pathological speech intelligibility by extracting acoustic features that are believed to be correlated with intelligibility. In [5]–[9], individual acoustic features such as jitter, shimmer, fundamental frequency, formant frequencies, or low-to-high modulation energy ratio (LHMR), are directly used to assess pathological speech intelligibility. In [5]–[7], [9], multiple acoustic features are combined through feature selection/reduction methods and regression models. Non-blind approaches rely on intelligible speech recordings from healthy speakers to estimate pathological speech intelligibility. In these approaches, healthy speech recordings are exploited in different manners. In [10], a speaker-independent Gaussian mixture model (GMM) is trained on healthy speech to create an intelligible reference model. By adapting the parameters of this reference model, a GMM-based supervector is created to represent the pathological speech signal. The intelligibility score is then obtained by training a regression model on the GMM-based supervector. A very similar approach is followed in [11]–[14], with the difference consisting in using an iVector or Gaussian posteriogram representation instead of a GMM-based supervector. In other non-blind approaches, pathological speech intelligibility is evaluated by training regression models on features produced by automatic speech recognition (ASR) systems, automatic speech alignment (ASA) systems, or phonological feature (PLF) extractor systems [15]–[22]. Commonly used features from such systems are the word error rate (WER), log-likelihood ratio, phoneme posteriors, and phonological features. These systems are typically trained using a large number of transcribed/segmented healthy speech recordings [15]–[22].

Although promising results have been shown using the above-mentioned approaches, several drawbacks arise when using them in practical scenarios. Most approaches require a large number of features for intelligibility prediction, increasing as a result the risk of over-fitting and limiting the performance in unseen data. In addition, non-blind approaches

P. Janbakhshi and H. Bourlard are with the Idiap Research Institute, Martigny 1920, Switzerland and École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland (e-mail: parvaneh.janbakhshi@idiap.ch, herve.bourlard@idiap.ch).

I. Kodrasi is with the Idiap Research Institute, Martigny 1920, Switzerland (email: ina.kodrasi@idiap.ch).

are typically complex and require a large number of healthy speech recordings for training, which might be infeasible for low-resource languages. Finally, non-blind approaches relying on ASR, ASA, and PLF systems require transcriptions of healthy and/or of pathological speech signals, which can be a time- and resource-consuming task.

To tackle the aforementioned drawbacks of state-of-the-art techniques, we have recently proposed a non-blind pathological speech intelligibility measure based on the extended short-time objective intelligibility (P-ESTOI) [23], [24]. P-ESTOI does not rely on a large number of features, does not require any training or a large number of healthy speech recordings, and has been shown to be highly correlated with subjective intelligibility ratings for patients suffering from several pathologies. However, for assessing the intelligibility of a sample utterance from a patient in P-ESTOI, recordings of the same utterance from several healthy speakers are needed such that an utterance-dependent reference model can be created. Intelligibility is then assessed through time-alignment of the pathological utterance with the utterance-dependent reference model. Consequently, P-ESTOI cannot be used in scenarios where such healthy recordings perfectly matching the phonetic content of the pathological speech signal are not available.

Aiming to develop an automatic intelligibility measure which does not require any time-alignment and can be used in phonetically unbalanced scenarios, in this paper we propose a subspace-based intelligibility (SBI) measure. This measure is inspired by the knowledge that speech pathologies typically decrease the degree of spectral modulation in pathological speech signals [25]. We hypothesize that pathology-induced spectral modulation changes are reflected in the subspace spanned by the most dominant speech spectral basis vectors. In addition, we hypothesize that the divergence between intelligible and pathological speech spectral subspaces (computed from healthy and pathological speech recordings that do not necessarily share the same phonetic content) can be used as an automatic intelligibility measure. Some encouraging preliminary results on the SBI measure have been presented in [26].

In comparison to [26], the contribution of this paper is four-fold. First, we propose to characterize spectral subspaces using a (possibly) different number of spectral basis vectors for the healthy and pathological speech signals. Second, we provide empirical evidence on i) the relation between the SBI measure and low-frequency components of the spectral modulation of speech which have been shown to be crucial for speech intelligibility, ii) the robustness of the SBI measure to gender variations, and iii) the robustness of the SBI measure to age variations. Third, we propose two techniques to incorporate short-time temporal information in the SBI measure. Finally, we provide an extensive experimental evaluation of the proposed measures to investigate their applicability in phonetically balanced and unbalanced scenarios and their generalisability across languages, i.e., English and Dutch, and across pathologies, i.e., Cerebral Palsy (CP) and hearing impairment (HI). Experimental results show that the proposed measures yield high and significant correlations with subjective intelligibility scores, while not requiring any training or a large number of healthy speech recordings and being applicable to phonetically unbalanced scenarios.

This paper is organized as follows. Section II presents a brief overview of the psychoacoustic evidence on the importance of spectral modulation frequencies on speech intelligibility. Section III describes the proposed SBI measure, and Section IV describes the proposed temporal extensions. Section V presents empirical insights on the relation between the proposed SBI measure and spectral modulation frequencies. In addition, its robustness to gender and age variations is empirically analyzed. Finally, experimental results using the proposed SBI measure and its temporal extensions are presented in Section VI.

II. Modulation spectrum and speech intelligibility

In this section, we present a brief overview of the psychoacoustic evidence supporting the relation between spectral modulation frequencies and speech intelligibility.

Fluctuations of the speech power spectrogram in time (at any frequency) and in frequency (at any time frame) are referred to as temporal and spectral modulations. Psychoacoustic studies have shown that the temporal and spectral modulations of speech are critical to speech perception, since they represent phonological information such as syllable boundaries and formant information [27]–[31]. The importance of spectro-temporal modulations to speech intelligibility is further confirmed by the success of several objective intelligibility measures typically used in speech enhancement which aim to incorporate (or indirectly assess) modulation cues, such as the speech transmission index [32], the spectro-temporal modulation index [33], LHMR [7], envelope power spectrum-based measures [34], [35], and the extended short-time objective intelligibility (ESTOI) measure [24]. Since our previously proposed pathological intelligibility measure P-ESTOI is based on ESTOI, it can be easily deduced that P-ESTOI also reflects differences in the spectro-temporal modulation of intelligible and pathological speech. While temporal modulations are indeed very important to speech intelligibility [28]–[31], the objective in this paper is to develop a measure which does not require time-alignment and which can be used in phonetically unbalanced scenarios. Hence, the proposed SBI measure can only reflect spectral modulation differences between healthy and pathological speech.

The effect of spectral modulation cues on the perceived speech intelligibility by human listeners (i.e., subjective speech intelligibility) has been extensively analyzed in [28]. In [28], the spectral modulation pattern is obtained by computing the Fourier transform of each time frame of the time-frequency (TF) representation of utterances. TF representations with a linear frequency axis result in spectral modulations in units of cycle/kHz, whereas TF representations with one-third octave band frequency axis result in spectral modulation in units of cycle/$\frac{1}{3}$octave. To investigate the spectral modulation frequencies contributing to speech intelligibility, the spectral modulation spectrum at each time frame is low-pass filtered at different cut-off frequencies. Using such low-pass filtering,

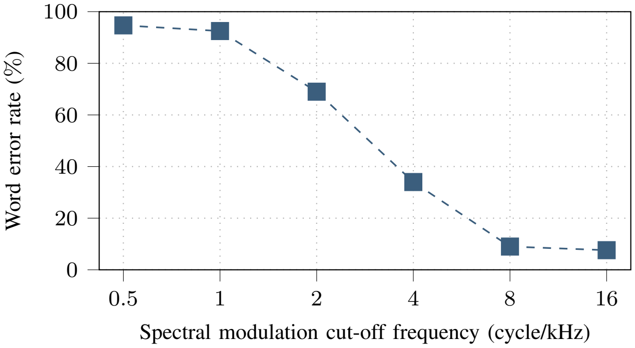

Fig. 1. Subjective intelligibility of low-pass spectral modulation filtered utterances based on the percentage of words misunderstood by human listeners. The spectral modulation spectrum of utterances is low-pass filtered at different cut-off frequencies (figure adapted from [28]).

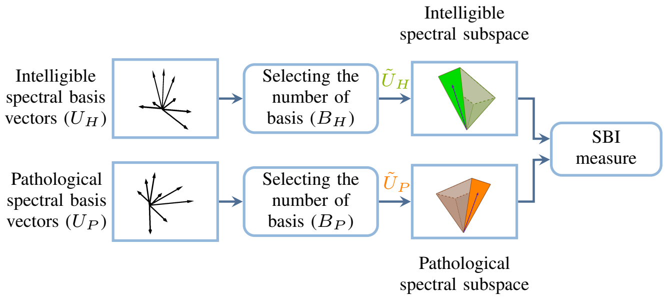
Intelligible spectral subspace
Intelligible spectral basis vectors (U_H)
Selecting the number of basis (B_H)
$\tilde{U}_H$
SBI measure
Pathological spectral basis vectors (U_P)
Selecting the number of basis (B_P)
$\tilde{U}_P$
Pathological spectral subspace

Fig. 2. Schematic representation of the proposed subspace-based intelligibility measure. Intelligible and pathological spectral basis vectors are obtained from intelligible (i.e., healthy) and pathological utterances. Low-dimensional spectral subspaces spanned by the most dominant intelligible and pathological spectral basis are created, where the number of dominant spectral basis vectors are automatically found. The pathological intelligibility score is computed as the distance between intelligible and pathological spectral subspaces.

the oscillations in the spectral modulation domain with frequencies above the considered cut-off frequency (i.e., higher-frequency components of the spectral modulation) are removed while oscillations below the considered cut-off frequency (i.e., lower-frequency components of the spectral modulation) are preserved. The time-domain signal corresponding to the low-pass filtered signal in the spectral modulation domain is reconstructed, and human listeners are asked to rate the intelligibility of these synthetically manipulated utterances.

Fig. 1 shows the effect of low-pass modulation filtering at different cut-off frequencies on the word error rate, i.e., the percentage of words misunderstood by listeners. As it can be observed, the word error rate significantly increases, i.e., speech intelligibility significantly decreases, when low spectral modulation frequencies are missing from the speech signal [28]. Low spectral modulation frequencies represent spectral amplitude fluctuations imposed by the vocal tract, i.e., formants and formant transitions [28]. Hence, it is to be expected that the removal of low spectral modulation frequencies yields a decrease in speech intelligibility. As will be shown in Section V-A, the SBI measure proposed in this paper responds to missing spectral modulation frequencies similarly to Fig. 1, i.e., similarly to how humans rate the perceived speech intelligibility when spectral modulation frequencies are missing in the speech signal.

III. Subspace-Based Pathological Speech Intelligibility Assessment

It is commonly accepted that speech spectrograms can be well approximated by low-rank matrices constructed using low-dimensional spectral patterns. Because of the reduced extent of articulatory movements in pathological speakers, the spectral variations in pathological speech are reduced [25]. Therefore, it can be expected that the dominant spectral patterns characterizing intelligible (healthy) speech differ from the ones characterizing pathological speech. Hence, we propose to estimate speech intelligibility by quantifying the distance between the spectral subspaces spanned by the dominant spectral basis of pathological speech and the dominant spectral basis of healthy speech. A schematic representation of the proposed SBI measure is depicted in Fig. 2. As depicted in this figure, SBI relies on i) computing spectral basis vectors characterizing spectral patterns in intelligible (i.e., healthy) utterances (referred to as intelligible spectral basis vectors), ii) computing spectral basis vectors characterizing spectral patterns in the test (i.e., pathological) utterance (referred to as pathological spectral basis vectors), iii) automatic selection of the number of spectral basis vectors used to create low-dimensional spectral subspaces corresponding to intelligible and pathological spectral patterns, and iv) computing the intelligibility score as the distance between the intelligible and pathological spectral subspaces. In the remainder of this section, the computational details of the proposed SBI measure are presented.

A. Computing intelligible spectral basis

While several techniques can be used to compute spectral basis such as approximate joint diagonalization [36], non-negative matrix factorization [37], and sparse coding [38], in this paper we propose to use the simple low-rank matrix decomposition minimizing the approximation error in the least-squares sense, i.e., the singular value decomposition (SVD). The SVD provides an analytical solution and results in a high performance for our application. To obtain meaningful spectral basis vectors, multiple utterances by several healthy speakers should be taken into account, such that the spectral basis vectors can capture patterns which are specific to intelligible speech but are independent of the particular speaker.

To obtain a signal representation resembling the transform properties of the auditory system, signals are first transformed to the TF domain by taking the logarithm of the one-third octave band spectrum [24], [28]. Let $H_{\text{stft}}(k,m)$ denote the short-time Fourier transform (STFT) of the speech signal $h(t)$, with k and m being the index of the frequency bin and of the time frame, respectively. The logarithm of the one-third octave band representation is computed as

$$H(j,m) = \log_{10}\sqrt{\sum_{k\in \text{CB}_j} |H_{\text{stft}}(k,m)|^2}, \qquad (1)$$

where j denotes the one-third octave band index and CB_j denotes the indices of STFT coefficients corresponding to the j^{th} one-third octave band. The parameter settings used in this paper for computing the TF representations are given in Section VI-B.

Let $\mathbf{H}_s$ denote the $(J \times M_s)$–dimensional TF representation of an utterance from healthy speaker s, with J being the total number of one-third octave bands and M_s being the total number of time frames. We consider TF representations of (possibly but not necessarily the same) utterances from different healthy speakers by concatenating them into a $(J \times M)$–dimensional matrix $\mathbf{H}$, where $M = \sum_{s=1}^{S} M_s$, i.e.,

$$\mathbf{H} = [\mathbf{H}_1 \; \mathbf{H}_2 \; \ldots \; \mathbf{H}_S], \tag{2}$$

with S being the total number of available healthy speakers. The SVD of $\mathbf{H}$ is given by

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{3}$$

with $\mathbf{U}$ being the $(J \times J)$–dimensional orthonormal matrix of left singular vectors representing spectral basis vectors, $\mathbf{\Sigma}$ being the $(J \times M)$–dimensional diagonal matrix of singular values σ_i assumed to be sorted in descending order, and $\mathbf{V}$ being the $(M \times M)$–dimensional orthonormal matrix of right singular vectors. The $(J \times B_H)$–dimensional matrix of dominant intelligible spectral basis vectors $\tilde{\mathbf{U}}_H$ is then constructed from the first B_H (with $B_H < J$) spectral basis vectors in $\mathbf{U}$. The selection of the number of intelligible spectral basis vectors B_H is described in Section III-C.

It should be noted that prior to computing the SVD, the matrices $\mathbf{H}_s$, $s = 1, \ldots, S$, in (2) are mean-centered in each octave band and scaled by $\frac{1}{\sqrt{M_s}}$ to remove the bias introduced by the number of time frames. This way, using the SVD in (3) to compute the spectral basis vectors is equivalent to using principal component analysis (PCA) [39]. Although mean-centering the representations in the framework of SVD is optional, it has been shown that the non-zero mean vector across time biases the first spectral basis vector to its direction rather than to the direction with maximal variability of spectral information [40], [41].

B. Computing test spectral basis vectors

To be able to assess intelligibility, the test (i.e., pathological) spectral basis vectors also need to be computed. Let $\mathbf{P}_r$ denote the $(J \times M_r)$–dimensional TF representation of the test utterance from patient r, with M_r denoting the total number of time frames. Similarly to Section III-A, the SVD of $\mathbf{P}_r$ is computed and the $(J \times J)$–dimensional orthonormal matrix $\mathbf{U}_P$ containing all pathological spectral basis vectors is obtained. Extracting only the dominant B_P basis vectors (with $B_P < J$) from $\mathbf{U}_P$, the $(J \times B_P)$–dimensional matrix of test spectral basis vectors $\tilde{\mathbf{U}}_P$ is constructed. The selection of the number of test spectral basis vectors B_P is described in Section III-C. It should be noted that differently from [26], to be able to obtain a better approximation of the intelligible and test representations, we allow the number of dominant spectral basis vectors for intelligible and test speech to be different, i.e., $B_H \neq B_P$.

C. Automatic selection of the number of spectral basis vectors

The number of spectral basis vectors B_H and B_P are hyperparameters of the proposed technique which obviously impact its performance. Using a large number of spectral basis vectors yields a better approximation of the considered TF representations. However, such an approximation is likely to capture not only spectral patterns important to speech intelligibility (i.e., the spectral basis vectors corresponding to larger singular values), but also spectral patterns describing extraneous variations such as speaker variability or noise (i.e., the spectral basis vectors corresponding to smaller singular values). The optimal number of spectral basis vectors should be as small as possible while at the same time it should yield a small approximation error to the original TF representation. Due to this inherent trade-off, in the following we propose to automatically select the number of spectral basis vectors B_H and B_P by adapting the L-curve method from [42], which has been successfully used to automatically select optimal regularization parameters in regularized least-squares techniques [43].

To automatically select the number of spectral basis vectors, we propose to use a parametric plot of the approximation error of the original TF representation versus the number of spectral basis vectors. This plot typically has an L-shape, with the corner (i.e., point of maximum curvature) representing a good compromise between the minimization of the approximation error and keeping the number of spectral basis vectors as low as possible. It should be noted that B_H and B_P can also be selected based on a user-defined threshold on the approximation error (as is typically done when using PCA for dimensionality reduction). However, using such a technique requires the user to define a threshold, introducing an additional hyperparameter which needs to be tuned.

The rank-B_H approximation of the original healthy representation $\mathbf{H}$ is obtained using the truncated SVD, i.e.,

$$\hat{\mathbf{H}} = \tilde{\mathbf{U}}_H \tilde{\mathbf{\Sigma}}_H \tilde{\mathbf{V}}_H^T, \tag{4}$$

where $\tilde{\mathbf{\Sigma}}_H$ denotes the $(B_H \times B_H)$–dimensional diagonal matrix containing the first B_H singular values and $\tilde{\mathbf{V}}_H$ is the $(B_H \times M)$–dimensional matrix containing the truncated right singular vectors in $\mathbf{V}$. The approximation error $\epsilon(B_H)$ of the intelligible TF representation $\mathbf{H}$ for different number of basis vectors B_H can be computed as

$$\epsilon(B_H) = \left\| \mathbf{H} - \hat{\mathbf{H}} \right\|_F^2 = \sum_{i=B_H+1}^{J} {\sigma_i}^2, \tag{5}$$

with $\|\cdot\|_F$ denoting the matrix Frobenious norm and σ_i denoting the i^{th} singular value. The approximation error $\epsilon(B_P)$ of the pathological TF representation $\mathbf{P}_r$ for different number of basis vectors B_P can be computed similarly to (5). To automatically select the number of spectral basis vectors B_H and B_P, the parametric plots of $\epsilon(B_H)$ versus B_H and of $\epsilon(B_P)$ versus B_P are constructed. Using the triangle method [44], the corner points of these parametric plots are computed and used as the number of dominant spectral patterns spanning the intelligible and pathological subspaces.

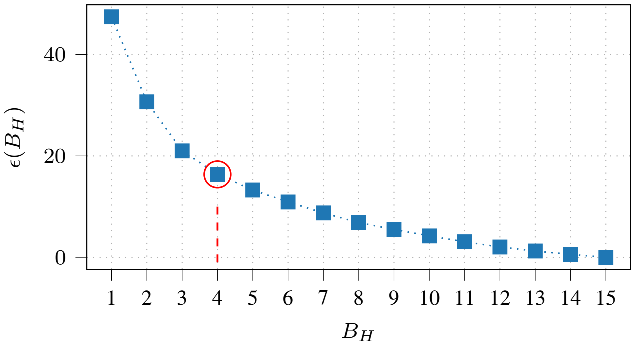

Fig. 3. Typical L-curve obtained for the approximation error $\epsilon(B_H)$ versus the number of basis vectors B_H for a sample utterance from the PC-GITA database [45]. The circle depicts the corner point automatically computed using the triangle method.

Fig. 3 depicts a typical parametric plot of $\epsilon(B_H)$ versus B_H for a sample utterance from the PC-GITA database [45]. As illustrated in this figure, this parametric plot has an L-shape, with the approximation error $\epsilon(B_H)$ decreasing as the number of spectral basis vectors B_H increases. The corner point automatically computed by the triangle method for this exemplary utterance is also depicted in this figure. Based on the L-curve criterion, using a larger number of basis vectors B_H than the one corresponding to the corner point (i.e., $B_H = 4$ in this example) does not provide any significant reduction in the approximation error. It should be noted that in this work, typical values for the number of basis vectors found with the L-curve method are 3 and 4.

D. Computing a distance measure between spectral basis vectors

As previously mentioned, the pathological intelligibility score is derived by quantifying the distance between the subspaces spanned by the spectral basis vectors in $\tilde{\mathbf{U}}_H$ (intelligible spectral subspace) and the spectral basis vectors in $\tilde{\mathbf{U}}_P$ (pathological spectral subspace). Since the dimensions of the intelligible and pathological subspaces are typically not the same, i.e., $B_P \neq B_H$, we use a distance measure between subspaces of different dimensions proposed in [46]. While other subspace distance measures can be used, in this work we use the Procrustes distance defined as

$$\delta(\tilde{\mathbf{U}}_H, \tilde{\mathbf{U}}_P) = 2\sqrt{\sum_{i=1}^{\min(B_H,B_P)} sin^2(\theta_i/2)}, \tag{6}$$

where θ_i denotes the i^{th} principal angle between subspaces which can be readily computed via the SVD[1] [46]. To be able to compare and combine intelligibility scores from different utterances (i.e., derived from using subspaces of different dimensions), the distance values are normalized to have a maximum value of 1 when the distance between the two subspaces is of the largest possible value, i.e., when $\theta_i = \pi/2,\ i = 1, ..., \min(B_H, B_P)$. Hence, the distance $\delta(\tilde{\mathbf{U}}_H, \tilde{\mathbf{U}}_P)$ obtained for each utterance is scaled by the factor

$$\alpha = \frac{1}{\sqrt{2\min(B_H, B_P)}}. \tag{7}$$

It should be noted that the proposed SBI measure is negatively associated with intelligibility, since the distance between pathological and intelligible spectral subspaces increases as pathological speech intelligibility decreases.

IV. Incorporating Temporal Information in SBI

The proposed SBI measure in Section III exploits only the spectral basis vectors in $\mathbf{U}_H$ and $\mathbf{U}_P$ for intelligibility assessment, while ignoring temporal patterns. Although temporal variations are important cues for speech intelligibility, the temporal basis of intelligible and pathological speech cannot be directly computed and compared to each other (because of unaligned and different phonetic contents in the TF representations of intelligible and pathological speech signals). In the following, we propose two viable approaches to incorporate short-time temporal information into the SBI measure. As will be shown in the experimental results in Section VI-C, using the proposed approaches to incorporate temporal information in the SBI measure can significantly improve the intelligibility assessment performance.

A. Dynamic SBI measure

Motivated by the dynamic PCA approach in [47], [48], in this section we propose to incorporate short-time temporal information into the SBI measure by modifying the TF representations through concatenating consecutive spectral vectors. Let $\mathbf{h}_m$ denote $(J \times 1)$–dimensional spectral vector at index m of the TF representation $\mathbf{H}$ (i.e., the m^{th} column of $\mathbf{H}$). By concatenating d such consecutive vectors, with d being a user-defined number ($d \ll M$, cf. Section VI-B), a new TF representation matrix $\mathbf{H}_{\text{DSBI}}$ is obtained, i.e.,

$$\mathbf{H}_{\text{DSBI}} = \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_{d+1} & \dots & \mathbf{h}_{(k-1)d+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{h}_d & \mathbf{h}_{2d} & \dots & \mathbf{h}_{kd} \end{bmatrix}, \tag{8}$$

where $k = \lfloor \frac{M}{d} \rfloor$. The matrix $\mathbf{H}_{\text{DSBI}}$ in (8) is a $(Jd \times k)$–dimensional matrix. The new pathological representation $\mathbf{P}_{\text{DSBI}}$ is obtained similarly to (8). Applying the same procedure as for computing the SBI measure in Section III to the modified representations $\mathbf{H}_{\text{DSBI}}$ and $\mathbf{P}_{\text{DSBI}}$, the dynamic SBI (DSBI) measure of pathological speech intelligibility is obtained. It should be noted that the modified TF representations $\mathbf{H}_{\text{DSBI}}$ and $\mathbf{P}_{\text{DSBI}}$ are of a larger spectral dimension than the original TF representations $\mathbf{H}$ and $\mathbf{P}$ (i.e., the number of rows in $\mathbf{H}_{\text{DSBI}}$ and $\mathbf{P}_{\text{DSBI}}$ is larger than the number of rows in $\mathbf{H}$ and $\mathbf{P}$). Consequently, the number of spectral basis vectors required to span these TF representations is also larger.

[1] It should be noted that the SVD used in [46] for computing the principal angles θ_i is unrelated to the SVD in (3) representing the spectral basis vectors.

B. Moving average SBI measure

Motivated by the moving average PCA model in [47], in this section we propose to incorporate short-time temporal information into the SBI measure by modifying the TF representations through a moving average model. Exploiting a moving average model serves to account for the short-time temporal correlation of speech signals, which is ignored in the SBI measure. It should be noted that while the DSBI measure proposed in Section IV-A considers multiple time frames simultaneously, the moving average SBI (MASBI) measure proposed in this section considers only a smoothed average across consecutive time frames. Unlike (8) where the spectral dimension is increased, the modified TF representation in MASBI has the original spectral dimension of (2).

The modified moving average TF representation is constructed as

$$\mathbf{H}_{\text{MASBI}} = [\mathbf{h}'_1 \quad \mathbf{h}'_2 \quad \dots \quad \mathbf{h}'_{M-q+1}], \tag{9}$$

where $\mathbf{h}'_m = \frac{1}{q} \sum_{j=m}^{m+q-1} \mathbf{h}_j$ for $m = 1, \dots, M-q+1$ and q is a user-defined number of time frames (cf. Section VI-B). The matrix $\mathbf{H}_{\text{MASBI}}$ in (9) is a $(J \times (M-q+1))$–dimensional matrix. The new pathological representation $\mathbf{P}_{\text{MASBI}}$ is also obtained similarly to (9). Applying the same procedure as for computing the SBI measure in Section III to the modified representations $\mathbf{H}_{\text{MASBI}}$ and $\mathbf{P}_{\text{MASBI}}$, the MASBI measure of pathological speech intelligibility is obtained.

V. Empirical insights into the proposed SBI measure

The objective of this section is to show through empirical analyses that the proposed SBI measure focuses on low-frequency spectral modulation cues to assess pathological speech intelligibility. This property can be justified by the psychoacoustic evidence confirming that low-frequency spectral modulations contribute to the perceived speech intelligibility by human listeners (cf. Section II). In addition, we provide empirical evidence on the robustness of SBI to gender and age variations. For these analyses, the algorithmic settings described in Section VI-B and recordings of healthy speakers from the PC-GITA database [45] are used. This database contains recordings of 50 Spanish-speaking healthy speakers (25 males and 25 females), with each speaker uttering 10 sentences. The age of the speakers ranges from 31 to 86 years old, with a median age of 62 years old [45].

A. SBI and spectral modulation of speech

In analogy to the experiment conducted in [28] (cf. Section II), in this section we analyze the effect of spectral modulation cues on the proposed SBI measure. To this end, the modulation spectrum obtained from the TF representation of each utterance from the PC-GITA database is low-pass filtered at different cut-off frequencies. Instead of asking human listeners to evaluate the perceived intelligibility of the low-pass spectral modulation filtered signals as in [28], we compute the proposed SBI measure based on the spectral basis vectors spanning the original utterances (representing e.g. healthy speech signals) and the low-pass filtered utterances (representing e.g. pathological speech signals).

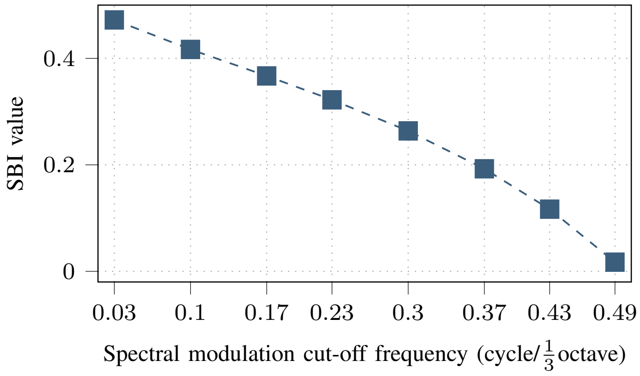

Fig. 4. Automatically estimated intelligibility using the proposed SBI measure for low-pass spectral modulation filtered utterances. The cut-off frequency units are cycle/$\frac{1}{3}$octave. The lack of low-frequency spectral modulations of speech has a similar effect on the estimated intelligibility by SBI as on the subjective intelligibility perceived by human listeners as depicted in Fig. 1.

Fig. 4 depicts the mean intelligibility estimated using the proposed SBI measure across all considered low-pass spectral modulation filtered utterances for different cut-off frequencies.[2] It can be observed that the effect of missing spectral modulation frequencies on SBI is similar to Fig. 1, i.e., similar to the effect of missing spectral modulation frequencies on the subjective speech intelligibility perceived by human listeners. In other words, the lack of low-frequency modulations in speech signals decreases the intelligibility estimated through the proposed SBI measure in a similar trend to how the perceived intelligibility by human listeners decreases. This observation shows that low-frequency components of spectral modulations are crucial for speech intelligibility assessment through SBI as they are also crucial for the perceived speech intelligibility by human listeners (cf. Section II). This observation is not surprising since the dominant spectral basis vectors obtained using the SVD span low-frequency spectral patterns. Consequently, the manipulation of these spectral patterns will be reflected in the proposed SBI measure.

B. Robustness of SBI to gender and age variations

A robust objective intelligibility measure should not be significantly impacted by non-pathological characteristics of speech such as gender- and age-related features. In this section we investigate the robustness of the proposed SBI measure to the gender and age of speakers. To ensure that the only source of variability is the gender or the age instead of pathology-related features, the following analyses are conducted on healthy (i.e., perfectly intelligible) speech recordings from the PC-GITA database.

To investigate the effect of gender on SBI, utterances of 20 (10 males and 10 females) speakers are used to represent

[2] It should be noted that the cut-off frequencies we use differ from [28] due to differences in the parameters of the TF representations. In addition, while [28] uses a linear frequency representation resulting in units of cycle/kHz, we use a logarithmic frequency representation resulting in units of cycle/$\frac{1}{3}$octave

the intelligible speech signals. To represent the test speech signals, utterances of 30 (15 males and 15 females) speakers are used. The disjoint subsets of intelligible and test speakers are randomly chosen from all available healthy speakers in the PC-GITA database, and the selection of these subsets is repeated 100 times. The SBI measure is then computed for each test utterance from each of the test male and female speakers. For each test utterance, the healthy TF representation is computed as in (2) by concatenating multiple instances of this utterance from the 20 speakers representing the intelligible speakers.

To investigate the robustness of SBI to age, a similar analysis is conducted by dividing the speakers into two age groups (i.e., a young group of speakers with age $\leqslant$ 62 years old and an old group of speakers with age $>$ 62 years old). To represent the intelligible speech signals, utterances of 18 (9 old and 9 young) speakers are used. To represent the test speech signals, utterances of 30 (15 old and 15 young) speakers are used. The disjoint subsets of intelligible and test speakers are also randomly chosen from all available healthy speakers in the PC-GITA database, and the selection of these subsets is also repeated 100 times. The SBI measure is then computed for each test utterance from each of the test young and old speakers. For each test utterance, the healthy TF representation is computed as in (2) by concatenating multiple instances of this utterance from the 18 speakers representing the intelligible speakers.

Figs. 5 and 6 depict the mean and standard deviation of the obtained SBI values for each utterance across the male and female speakers and across the young and old speakers. These results are obtained for one disjoint subset of intelligible and test speakers randomly chosen from all available healthy speakers in the PC-GITA database. It can be observed that the obtained mean SBI values are very similar across the two gender and age groups, independently of the considered utterance. This shows that the proposed SBI measure is barely affected by the gender or age of speakers. In addition, it can be observed that the mean SBI values for all groups of speakers and for all utterances are typically low. This is to be expected, since the test signals are perfectly intelligible independently of the gender or age of speakers and the distance between spectral subspaces of intelligible speech should be minimal.

To evaluate whether there are significant differences between the mean SBI values for each utterance across the groups of speakers (i.e., male vs. female groups and old vs. young groups), an independent samples t-test is conducted. The t-test is conducted for each repetition of the speakers' subset selection in both gender- and age-based analyses. Out of the 10 considered utterances, the average number of utterances across all repetitions which yield a statistically significant difference (i.e., $p < 0.01$) between the mean SBI values of male and female speakers is only 1. Similarly, the average number of utterances across all repetitions which yield a statistically significant difference (i.e., $p < 0.01$) between the mean SBI values of old and young speakers is also only 1. For the speakers' subset selection used in Fig. 5 and Fig. 6, no statistically significant differences for any of the utterances are found. Hence, it can be said that the difference in the mean SBI values across male and female speakers and the difference in the mean SBI values across old and young speakers is generally not statistically significant.

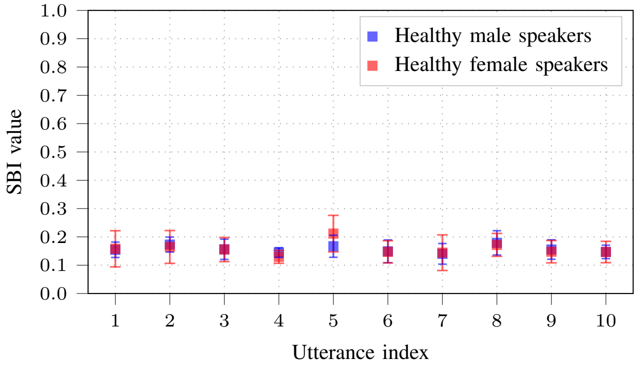

Fig. 5. Mean and standard deviation of the obtained SBI values across male and female speakers for one repetition of the speakers' subset selection.

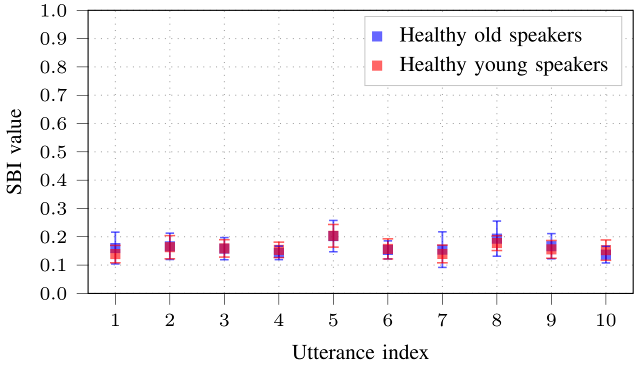

Fig. 6. Mean and standard deviation of the obtained SBI values across old and young speakers for one repetition of the speakers' subset selection.

In summary, our analyses show that the proposed SBI measure is not sensitive to the gender and age of speakers and is able to construct representations which can mainly reflect intelligibility-related degradations.

VI. Results And Discussion

In this section, the performance of the proposed intelligibility measures is extensively investigated and compared to state-of-the-art measures. To demonstrate the applicability of the proposed measures for several languages and pathologies, we evaluate the performance on databases of English-speaking CP patients and Dutch-speaking HI patients with a speech disorder. To demonstrate the applicability of the proposed measures for a wide range of scenarios, we consider both phonetically balanced and phonetically unbalanced scenarios.

A. Databases and preprocessing

The performance of the proposed intelligibility measures is evaluated on the following two pathological speech databases.

Universal access speech (UA-Speech) database [49]. The UA-Speech database includes recordings of 15 English-speaking dysarthric patients (11 males, 4 females) suffering from CP and of 13 healthy speakers (9 males, 4 females). Each speaker read 763 isolated words, with 155 of the words uttered three times and referred to as common words (CW). The remaining 298 words were uttered only once and are referred to as uncommon words (UW). A 7-channel microphone array is used for recording the speakers at 16 kHz sampling rate. For the evaluations presented in this paper, we consider the recordings of the 5th (arbitrarily selected) channel. In addition, to extract speech-only segments, an energy-based voice activity detection is applied to the speech recordings [50]. The subjective intelligibility scores of patients range from 2% to 95% [49].

Dutch corpus of pathological and normal speech (COPAS) [51]. In addition to the English-speaking CP patients, we consider recordings of 16 Dutch-speaking HI patients with a speech disorder (6 males, 10 females) and of 22 healthy speakers (11 males, 11 females). For each speaker, recordings of 10 sentences sampled at 16 kHz are used. Individual words are extracted from all sentences using forced alignment from an ASR system followed by manual corrections, resulting in 47 available words for each speaker. The subjective intelligibility scores of patients range from 53% to 98% [51].

B. Algorithmic settings, state-of-the-art measures, scenarios, and evaluation

In this section, we present the algorithmic settings for the implementation of the proposed intelligibility measures. In addition, the considered state-of-the-art measures that are compared to the proposed measures are briefly described. Finally, the considered scenarios and the performance evaluation metrics are presented.

Algorithmic settings. To compute intelligible representations for the proposed measures, we use speech signals from both healthy male and female speakers. Intelligible representations for the CP patients are constructed using the speech signals of 9 male and 4 female healthy speakers from the UA-Speech database. Intelligible representations for the HI patients are constructed using the speech signals of 11 male and 11 female healthy speakers from the COPAS database. To obtain the TF representations in (1), the same STFT and octave band settings as in [23] are used. The empirically selected number of time frames used to incorporate temporal information in the DSBI and MASBI measures is $d = 5$ (cf. (8)) and $q = 9$ (cf. (9)), respectively. It should be noted that the computation of spectral basis vectors for the proposed measures is efficient when PCA is used in practice. For the SBI and MASBI measures, spectral basis vectors are obtained by applying PCA on the 15×15–dimensional correlation matrices $\mathbf{H}\mathbf{H}^T$ and $\mathbf{H}_{\text{MASBI}}\mathbf{H}_{\text{MASBI}}^T$. Running PCA for such matrices on a computer with a 2.7 GHz processor and 8 GB RAM requires only 0.0003 seconds. For the DSBI measure, spectral basis vectors are obtained by applying PCA on the 75×75–dimensional matrix $\mathbf{H}_{\text{DSBI}}\mathbf{H}_{\text{DSBI}}^T$ (since $d = 5$). Running PCA for such matrices on a computer with a 2.7 GHz processor and 8 GB RAM requires only 0.003 seconds.

State-of-the-art measures. The performance of the proposed measures is compared to the performance of several non-blind state-of-the-art measures, i.e., P-ESTOI [23], iVector-based approach [13], and ASR-based approach [13]. The algorithmic settings for P-ESTOI are the same as in [23], and the performance of P-ESTOI is evaluated on both considered databases. For the iVector- and ASR-based approaches, we report the results from [13] where these approaches are evaluated only on the UA-Speech database following a leave-one-subject-out validation strategy.

Scenarios. The performance of the considered intelligibility measures is evaluated for the following two scenarios.

Phonetically balanced scenarios. In these scenarios, we assume that all speakers (healthy and pathological) utter exactly the same words. All 763 available words are considered for the UA-Speech database, and all 47 available words are considered for the COPAS database. The intelligibility score is calculated for each word, and the final intelligibility score is computed as the mean intelligibility score across all words. Only in such phonetically balanced scenarios can the performance of the proposed measures be compared to the performance of P-ESTOI (since otherwise healthy speech reference models for P-ESTOI cannot be constructed). In addition, the performance of the iVector- and ASR-based approaches in [13] has been analyzed also in such a phonetically balanced scenario (only for the UA-Speech database).

Phonetically unbalanced scenarios. In these scenarios, the applicability of the proposed measures is analyzed in the presence of phonetic variability in the considered speech signals from each speaker. Since speakers utter different words in such scenarios, a robust spectral subspace can only be constructed when longer utterances (i.e., longer than a single word) are taken into account. Different sets of words are concatenated to create longer utterances for each speaker, and a single intelligibility score is estimated for each patient. Since the UA-Speech database contains a large number of words which can be combined in different ways for different speakers, these analyses are done on the UA-Speech database. We assess the effect of different levels of phonetic variability on the proposed intelligibility measures by concatenating multiple words for each speaker in the following manners.

i) The phonetic content within the speakers in each group is the same, while the phonetic content across the two groups of speakers is partially different. To generate this scenario, the set UW is randomly divided into two subsets of equal size (149 words). The utterance uttered by all healthy speakers is created by concatenating one such subset of UW and one repetition (155 words) of the set CW. The utterance uttered by all pathological speakers is created by concatenating the other subset of UW and one repetition (155 words) of the set CW. The total number of concatenated words in each utterance is 304.

ii) The phonetic content within the speakers in each group is the same, while the phonetic content across the two groups of speakers is completely different. To generate this scenario, a similar procedure as in i) is followed. Differently from i), the set CW is also randomly divided into two disjoint subsets (of size 77 and 78 words). The utterance uttered by all healthy speakers is created by concatenating the previously considered subset of UW and one such subset of CW. The utterance uttered by all pathological speakers is created by concatenating the previously considered subset of UW and the other subset of CW. The total number of concatenated words for each healthy speaker is 226, whereas the total number of concatenated words for each pathological speaker is 227.

iii) The phonetic content across all speakers is partially different. To generate this scenario, the utterance for each speaker is created by concatenating 200 randomly selected words from the UW and CW sets. Since there are only a total of 763 words available, there is a partial overlap between the phonetic content across the different speakers.

iv) The phonetic content across all speakers is completely different. To generate this scenario, the utterance for each speaker is created by concatenating 16 distinct (and randomly selected) words from the UW and CW sets.

The subset of words to be concatenated for creating longer utterances for each speaker in the above-mentioned scenarios is randomly selected. This selection is repeated 100 times, and the performance of the proposed measures is analyzed in terms of the mean and standard deviation of the performance across all repetitions.

Performance metrics. To evaluate the performance of the automatic pathological intelligibility measures, the Pearson correlation coefficient (R) and the Spearman rank correlation coefficient (R_S) between the automatically estimated intelligibility and the subjective intelligibility scores of the CP patients [49] and HI patients [51] are computed. In addition, the statistical significance of these correlation values is also assessed. To evaluate the statistical significance, the critical values of R and R_S, denoted by R_c and R_{Sc}, respectively, are computed using a significance level $\alpha = 0.05$ and taking into account the number of patients in each database [52], [53]. The obtained critical values are presented in Table I. The correlation values obtained for the different intelligibility measures are considered to be statistically significant if $|R| \geq |R_c|$ and $|R_S| \geq |R_{Sc}|$.

C. Performance in phonetically balanced scenarios

In this section, the performance of the proposed measures in phonetically balanced scenarios is compared to the performance of state-the-art measures.

Table II presents the Pearson and Spearman correlation values obtained for the CP and HI patients using the proposed measures and the P-ESTOI measure. In addition, the Pearson correlation values obtained for the CP patients using the iVector- and ASR-based approaches in [13] are also presented. As previously mentioned, only the Pearson correlation coefficients for the CP patients have been reported in [13]. Hence, results for HI patients and Spearman correlation values for CP patients are not available. To assess the statistical significance of the reported correlation values, entries in Table II are compared to the corresponding critical correlation values in Table I (cf. Section VI-B).

TABLE I
CRITICAL VALUES FOR THE PEARSON AND SPEARMAN CORRELATION COEFFICIENTS OBTAINED USING $\alpha = 0.05$. THE NUMBER OF PAIRS OF SCORES IS CONSIDERED TO BE THE NUMBER OF PATIENTS IN EACH DATABASE [52], [53].

15 English CP patients		16 Dutch HI patients	
R_c	R_{Sc}	R_c	R_{Sc}
-0.441	-0.443	-0.426	-0.443

TABLE II
PERFORMANCE OF THE PHONETICALLY BALANCED INTELLIGIBILITY ASSESSMENT ON THE ENGLISH CP AND DUTCH HI DATABASES USING THE PROPOSED (I.E., SBI, DSBI, AND MASBI) AND STATE-OF-THE-ART (I.E., P-ESTOI, IVECTOR, AND ASR) MEASURES. THE ENTRY DENOTED BY $\{\cdot\}^*$ INDICATES NON-SIGNIFICANT CORRELATION, AND ENTRIES DENOTED BY {-} INDICATE THAT CORRELATION VALUES ARE NOT AVAILABLE.

	15 English CP patients		16 Dutch HI patients	
Measures	R	R_S	R	R_S
P-ESTOI	0.944	0.945	0.804	0.805
iVector	0.74	-	-	-
ASR	0.55	-	-	-
SBI	-0.856	-0.877	-0.480	-0.397^*
DSBI	-0.863	-0.934	-0.641	-0.603
MASBI	-0.821	-0.877	-0.682	-0.650

It can be observed that P-ESTOI gives the highest correlation values on both databases, which is to be expected since P-ESTOI takes both the temporal and spectral distortions into account by aligning the pathological speech signals to the intelligible reference representations. However, this limits the application of P-ESTOI to only such phonetically balanced scenarios. For the CP patients, the proposed SBI, DSBI, and MASBI measures also yield very high and significant correlations with the subjective intelligibility scores, significantly outperforming the state-of-the-art iVector- and ASR-based approaches. In comparison to the SBI measure, incorporating short-time temporal information as in the DSBI measure slightly increases the obtained correlation on this database. Incorporating short-time temporal information as in the MASBI measure slightly decreases the Pearson correlation coefficient, whereas the Spearman rank correlation coefficient is the same as for the SBI measure. However, the SBI measure does not show significant Spearman rank correlation values on the HI database. Incorporating short-time temporal information through the DSBI and MASBI measures significantly improves the performance over the SBI measure on this database.

It should be noted that the results presented here are obtained using an arbitrarily selected subset of healthy speakers to generate intelligible representations. We have additionally investigated the sensitivity of the proposed measures to the choice of healthy speakers for computing intelligible representations. Although we have omitted these results due to space constraints, they show that the performance of the proposed measures is insensitive to the specific healthy speakers used for generating reference representations. Additionally, we have compared the proposed measures to other state-of-the-art blind intelligibility measures which have been shown to yield a high correlation with subjective intelligibility scores in [7], i.e., kurtosis of the linear prediction residual, voicing percentage, LHMR, etc. [7]. However, these measures resulted in a significantly worse performance than the proposed measures, and these results are omitted in this paper for the sake of brevity.

In summary, it can be said that the proposed measures are applicable to phonetically balanced scenarios and result in high and significant correlations with subjective intelligibility scores. In addition, it can be said that incorporating short-time temporal information (i.e., as in the DSBI and MASBI measures) can yield a significant performance improvement as opposed to considering only spectral information (i.e., as in the SBI measure).

D. Performance in phonetically unbalanced scenarios

In this section, the performance of the proposed measures is analyzed in phonetically unbalanced scenarios. It should be noted that the P-ESTOI measure is inapplicable to such scenarios since the phonetic content among all speakers should be the same to be able to create an intelligible reference representation.

Table III presents the mean and standard deviation of the Pearson and Spearman rank correlation values across all repetitions of words' subset selection obtained using the proposed SBI, DSBI, and MASBI measures for all considered phonetically unbalanced scenarios. To assess the statistical significance of the reported correlation values, entries in Table III are compared to the corresponding critical correlation values in Table I (cf. Section VI-B). Overall it can be observed that all proposed measures typically yield high and significant correlations with the subjective intelligibility scores. In addition, the performance of individual measures for scenarios i)–iii) is very similar, showing that the different levels of phonetic variability in these scenarios do not significantly affect the performance of the proposed measures. However, it can be observed that the performance of the proposed measures for scenario iv) is lower than for the other scenarios. This performance degradation in scenario iv) is to be expected since intelligibility is assessed using only 16 words which are different across all speakers. Such a small number of words with different phonetic content does not suffice to construct a robust subspace reflecting speech intelligibility. While the performance of all proposed measures decreases in this scenario, the performance of the proposed DSBI measure is particularly lower. The DSBI measure relies on a TF representation of a larger spectral dimension than the SBI and MASBI measures. The number of spectral basis vectors required to span the intelligible and test representations for this measure is larger. Consequently, to construct robust subspaces when the phonetic content among speakers differ, longer utterances are required for this measure than for the SBI and MASBI measures.

In summary, it can be said that the proposed measures are applicable to phonetically unbalanced scenarios and result in high and significant correlations with subjective intelligibility scores. Since the phonetic content across speakers differs in such scenarios, incorporating short-time temporal information (i.e., as in the DSBI and MASBI measures) does not yield a performance improvement as opposed to considering only spectral information (i.e., as in the SBI measure).

TABLE III
PERFORMANCE OF THE PHONETICALLY UNBALANCED INTELLIGIBILITY ASSESSMENT ON THE ENGLISH CP DATABASE USING THE PROPOSED MEASURES. THE ENTRIES DENOTED BY $\{\cdot\}^*$ INDICATE NON-SIGNIFICANT CORRELATIONS.

Measures	R	R_S
Phonetically unbalanced scenario i)		
SBI	-0.742 ± 0.025	-0.760 ± 0.033
DSBI	-0.693 ± 0.050	-0.714 ± 0.060
MASBI	-0.726 ± 0.040	-0.763 ± 0.057
Phonetically unbalanced scenario ii)		
SBI	-0.735 ± 0.028	-0.755 ± 0.038
DSBI	-0.699 ± 0.059	-0.731 ± 0.062
MASBI	-0.710 ± 0.060	-0.739 ± 0.073
Phonetically unbalanced scenario iii)		
SBI	-0.733 ± 0.033	-0.758 ± 0.041
DSBI	-0.690 ± 0.062	-0.718 ± 0.074
MASBI	-0.721 ± 0.052	-0.755 ± 0.061
Phonetically unbalanced scenario iv)		
SBI	-0.697 ± 0.070	-0.724 ± 0.077
DSBI	$-0.372^{*} \pm 0.157$	$-0.407^{*} \pm 0.166$
MASBI	-0.651 ± 0.112	-0.653 ± 0.122

The presented analyses show the successful applicability of the proposed measures on speech disorders arising due to CP and HI. The applicability of the proposed measures on other types of speech disorders should be further investigated. To the best of our knowledge, a systematic comparison of spectral modulation changes across different pathologies has never been done. If the induced spectral modulation changes are dependent on the pathology, it can be expected that the proposed measures perform differently on different pathologies. Given the advantageous intelligibility assessment results, such spectral subspace-based representation of speech might prove useful in other applications of pathological speech assessment, e.g., in pathological speech detection. This research direction remains to be investigated in the future.

VII. CONCLUSION

In this paper, we have proposed the automatic pathological speech intelligibility SBI measure, which is based on the assessment of the distance between subspaces spanned by dominant spectral patterns of intelligible (i.e., healthy) and pathological speech. Exploiting psychoacoustic evidence on the importance of spectral modulation cues to the perceived speech intelligibility, we have shown that the proposed SBI measure is advantageous since it can capture pathology-induced distortions in the spectral modulation cues. In addition, we have shown that the proposed measure is robust to gender- and age-induced changes in the acoustical properties of signals. To be able to additionally track possible degradations in the temporal structure of the pathological speech signal, we have also proposed two extensions of the SBI measure, i.e., the DSBI and MASBI measures. Experimental results show that the proposed measures obtain high correlations with subjective intelligibility scores, with the incorporation of temporal information into the DSBI and MASBI measures yielding a better performance in phonetically balanced scenarios. In addition, it has been shown that the proposed measures outperform several non-blind state-of-the-art measures, while not requiring any regression training or a large amount of healthy speech training data and being also applicable to phonetically unbalanced scenarios.

Acknowledgment

The authors would like to acknowledge the support of the Swiss National Science Foundation project no CRSII5_173711 "MoSpeeDi" on "*Motor Speech Disorders: characterizing phonetic speech planning and motor speech programming/execution and their impairments*". They would also like to thank the project partners from Geneva University, University Sorbonne Nouvelle, and Geneva University Hospitals for a fruitful collaboration.

References

[1] J. E. Sussman and K. Tjaden, "Perceptual measures of speech from individuals with Parkinson's disease and multiple sclerosis: intelligibility and beyond," *Journal of Speech and Hearing Research*, vol. 55, no. 4, pp. 1208–1219, Aug. 2012.

[2] N. Miller, "Measuring up to speech intelligibility," *International Journal of Language & Communication Disorders*, vol. 48, no. 6, pp. 601–612, Nov.-Dec. 2013.

[3] J. Oates, "Auditory-perceptual evaluation of disordered voice quality: pros, cons and future directions," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 1, pp. 49–56, Feb. 2009.

[4] L. Baghai-Ravary and S. Beet, *Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders*. New York, USA: Springer, 2012.

[5] M. S. Paja and T. H. Falk, "Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech," in *Proc. 13th Annual Conference of the International Speech Communication Association*, Oregon, USA, Sep. 2012, pp. 62–65.

[6] R. Hummel, W. Y. Chan, and T. H. Falk, "Spectral features for automatic blind intelligibility estimation of spastic dysarthric speech," in *Proc. 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, Aug. 2011, pp. 3017–3020.

[7] T. H. Falk, W. Y. Chan, and F. Shein, "Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility," *Speech Communication*, vol. 54, no. 5, pp. 622–631, June 2012.

[8] T. Haderlein, A. Schützenberger, M. Döllinger, and E. Noeth, "Robust automatic evaluation of intelligibility in voice rehabilitation using prosodic analysis," in *Proc. 20th International Conference on Text, Speech, and Dialogue*, Prague, Czech Republic, Aug. 2017, pp. 11–19.

[9] A. R. Fletcher, A. A. Wisler, M. J. McAuliffe, K. L. Lansford, and J. M. Liss, "Predicting intelligibility gains in dysarthria through automated speech feature analysis," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 11, pp. 3058–3068, Nov. 2017.

[10] T. Bocklet, K. Riedhammer, U. Eysholdt, and T. Haderlein, "Automatic intelligibility assessment of speakers after laryngeal cancer by means of acoustic modeling," *Journal of Voice*, vol. 26, no. 3, pp. 390–397, May 2012.

[11] D. Martínez, P. Green, and H. Christensen, "Dysarthria intelligibility assessment in a factor analysis total variability space," in *Proc. 14th Annual Conference of the International Speech Communication Association*, Lyon, France, Aug. 2013, pp. 2133–2137.

[12] L. Imed, B. K. Waad, F. Corinne, and M. Christine, "Automatic prediction of speech evaluation metrics for dysarthric speech," in *Proc. 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, Aug. 2017, pp. 1834–1838.

[13] D. Martínez, E. Lleida, P. Green, H. Christensen, A. Ortega, and A. Miguel, "Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace," *ACM Transactions on Accessible Computing*, vol. 6, no. 3, pp. 1–21, June 2015.

[14] S. Kalita, S. R. Mahadeva Prasanna, and S. Dandapat, "Intelligibility assessment of cleft lip and palate speech using Gaussian posteriograms based on joint spectro-temporal features," *Journal of the Acoustical Society of America*, vol. 144, no. 4, pp. 2413–2423, Oct. 2018.

[15] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "PEAKS - A system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, no. 5, pp. 425–437, Jan. 2009.

[16] C. Middag, G. V. Nuffelen, J.-P. Martens, and M. De Bodt, "Objective intelligibility assessment of pathological speakers," in *Proc. 9th Annual Conference of the International Speech Communication Association*, Brisbane, Australia, Sep. 2008, pp. 1745–1748.

[17] C. Middag, Y. Saeys, and J.-P. Martens, "Towards an ASR-free objective analysis of pathological speech," in *Proc. 11th Annual Conference of the International Speech Communication Association*, Makuhari, Japan, Sep. 2010, pp. 294–297.

[18] C. Middag, J.-P. Martens, G. V. Nuffelen, and M. De Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 1–9, May 2009.

[19] G. V. Nuffelen, C. Middag, M. De Bodt, and J.-P. Martens, "Speech technology-based assessment of phoneme intelligibility in dysarthria," *International Journal of Language & Communication Disorders*, vol. 44, no. 5, pp. 716–730, Sep. 2009.

[20] T. Haderlein, S. Steidl, E. Nöth, F. Rosanowski, and M. Schuster, "Automatic recognition and evaluation of tracheoesophageal speech," in *Proc. 7th International Conference on Text, Speech and Dialogue*, Brno, Czech Republic, Sep. 2004, pp. 331–338.

[21] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, and F. Girardi, "Assessment of speech intelligibility in parkinson's disease using a speech-to-text system," *IEEE Access*, vol. 5, pp. 22 199–22 208, Oct. 2017.

[22] M. J. Kim, Y. Kim, and H. Kim, "Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 694–704, Apr. 2015.

[23] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Pathological speech intelligibility assessment based on the short-time objective intelligibility measure," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brighton, UK, May 2019, pp. 6405–6409.

[24] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, Aug. 2016.

[25] K. M. Rosen, R. D. Kent, A. L. Delaney, and J. R. Duffy, "Parametric quantitative acoustic analysis of conversation produced by speakers with dysarthria and healthy speakers," *Journal of Speech Language and Hearing Research*, vol. 49, no. 2, pp. 395–411, Apr. 2006.

[26] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Spectral subspace analysis for automatic assessment of pathological speech intelligibility," in *Proc. 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep. 2019, pp. 3038–3042.

[27] H. Hermansky, "Speech recognition from spectral dynamics," *Sadhana*, vol. 36, no. 5, pp. 729–744, Oct. 2011.

[28] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLOS Computational Biology*, vol. 5, no. 3, pp. 1–14, Mar. 2009.

[29] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, Oct. 1995.

[30] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, Oct. 1994.

[31] F.-G. Zeng, K. Nie, G. S. Stickney, Y.-Y. Kong, M. Vongphoe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," *Proceedings of the National Academy of Sciences*, vol. 102, no. 7, pp. 2293–2298, Feb. 2005.

[32] H. J. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *The Journal of the Acoustical Society of America*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[33] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Communication*, vol. 41, no. 2, pp. 331–348, Oct. 2003.

[34] T. Biberger and S. D. Ewert, "The role of short-time intensity and envelope power for speech intelligibility and psychoacoustic masking," *Journal of the Acoustical Society of America*, vol. 142, no. 2, pp. 1098–1111, Aug. 2017.

[35] S. Jorgensen, S. D. Ewert, and T. Dau, "A multi-resolution envelope-power based model for speech intelligibility," *Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 436–446, July 2013.

[36] J. Cardoso and A. Souloumiac, "Jacobi angles for simultaneous diagonalization," *SIAM Journal on Matrix Analysis and Applications*, vol. 17, no. 1, pp. 161–164, Jan. 1996.

[37] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[38] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, June 1996.

[39] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, *A Practical Approach to Microarray Data Analysis*. Boston, MA: Springer US, 2003, ch. Singular Value Decomposition and Principal Component Analysis, pp. 91–109.

[40] J. Cadima and I. T. Jolliffe, "On relationships between uncentred and column-centred principal component analysis," *Pakistan Journal of Statistics*, vol. 25, no. 4, pp. 473–503, Oct. 2009.

[41] N. Alexandris, S. Gupta, and N. Koutsias, "Remote sensing of burned areas via PCA, part 1; centering, scaling and EVD vs SVD," *Open Geospatial Data, Software and Standards*, vol. 2, no. 1, pp. 1–11, July 2017.

[42] P. C. Hansen, "Analysis of discrete ill-posed problems by means of the L-curve," *SIAM Review*, vol. 34, no. 4, pp. 561–580, Dec. 1992.

[43] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1879–1890, Sep. 2013.

[44] J. Castellanos, S. Gómez, and V. Guerra, "The triangle method for finding the corner of the L-curve," *Applied Numerical Mathematics*, vol. 43, no. 4, pp. 359–373, Dec. 2002.

[45] J. R. Orozco, J. D. Arias-Londoño, J. Vargas-Bonilla, M. González-Rátiva, and E. Noeth, "New spanish speech corpus database for the analysis of people suffering from parkinson's disease," in *Proc. 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, May 2014, pp. 342–347.

[46] K. Ye and L. H. Lim, "Schubert varieties and distances between subspaces of different dimensions," *SIAM Journal on Matrix Analysis and Applications*, vol. 37, no. 3, pp. 1176–1197, Jan. 2016.

[47] Z. Zhao and F. Liu, "Industrial monitoring based on moving average PCA and neural network," in *Proc. 30th Annual Conference of IEEE Industrial Electronics Society*, Busan, South Korea, Nov. 2004, pp. 2168–2171.

[48] W. Ku, R. H. Storer, and C. Georgakis, "Disturbance detection and isolation by dynamic principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 30, no. 1, pp. 179–196, Nov. 1995.

[49] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proc. 9th Annual Conference of the International Speech Communication Association*, Brisbane, Australia, Sep. 2008, pp. 1741–1744.

[50] P. Boersma and D. Weenink, "PRAAT, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9, pp. 341–345, Jan. 2002.

[51] G. V. Nuffelen, M. De Bodt, C. Middag, and J.-P. Martens, "Dutch corpus of pathological and normal speech (COPAS)," Antwerp University Hospital and Ghent University, Belgium, Tech. Rep., 2009.

[52] D. Zwillinger and S. Kokoska, *CRC Standard Probability and Statistics Tables and Formula*. New York: Chapman and Hall, 2000, ch. Nonparametric Statistics.

[53] ——, *CRC Standard Probability and Statistics Tables and Formula*. New York: Chapman and Hall, 2000, ch. Standard Normal Distribution.

Parvaneh Janbakhshi (S'19) received the Master of Science (M.S.) in bioelectrical engineering from Sharif University of Technology, Iran, in 2016. She is currently a research assistant at Idiap Research Institute, Martigny, Switzerland and pursuing a Ph.D. in Electrical Engineering at the Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland. Her Ph.D. thesis research is focusing on the automatic assessment of pathological speech. Her research interest includes pathological speech and audio signal processing and machine learning.

Ina Kodrasi (S'11-M'16) received the Master of Science degree in Communications, Systems, and Electronics from Jacobs University Bremen, Bremen, Germany, in 2010, and the Ph.D. degree from the University of Oldenburg, Oldenburg, Germany, in 2015. She was a research associate and a postdoctoral researcher at the Signal Processing Group, University of Oldenburg from 2010–2015 and 2015–2017 respectively, where she worked on speech enhancement. From 2010 to 2011, she was also with the Project Group Hearing, Speech and Audio Technology, Fraunhofer Institute for Digital Media Technology, Oldenburg, Germany, where she worked on microphone array beamforming. Since December 2017, she has been a postdoctoral researcher with the Idiap Research Institute, Martigny, Switzerland, working in the field of speech signal processing for clinical applications. She is an elected member of the IEEE Technical Committee of Audio and Acoustic Signal Processing and of the EURASIP Special Area Team of Acoustic, Speech and Music Signal Processing.

Hervé Bourlard received the Electrical and Computer Science Engineering degree and the Ph.D. degree in Applied Sciences both from "Faculté Poly technique de Mons", Mons, Belgium. Starting his research career as a member of the Scientific Staff at the Philips MBLE Research Laboratory of Brussels, he is now (since 1996) Director of the Idiap Research Institute and Full Professor at the Swiss Federal Institute of Technology Lausanne (EPFL). He was also the Founding Director of the Swiss NSF National Centre of Competence in Research on "Interactive Multimodal Information Management (2001–2013)". Having spent (since 1988) several long-term and short-term visits at the International Computer Science Institute (ICSI), Berkeley, CA, he is now an ICSI External Fellow and Emeritus Trustee. His main research interests include statistical pattern classification, signal processing, multi-channel processing, artificial neural networks, and applied mathematics, with applications to a wide range of Information and Communication Technologies, including spoken language processing, speech and speaker recognition, language modeling, multimodal interaction, and augmented multi-party interaction. H. Bourlard is the author/co-author/editor of 9 books and over 350 reviewed (journal, conference, and book chapter) papers, including one IEEE journal paper award. He is an IEEE Fellow, an ISCA Fellow, a Senior Member of ACM, and an elected member of the Swiss Academy of Engineering Sciences. Having worked for academia as well as large and small (startup) industries, he is the recipient of several scientific and entrepreneurship awards.