INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Spectral Subspace Analysis for Automatic Assessment of Pathological Speech Intelligibility

Parvaneh Janbakhshi[1,2]*, Ina Kodrasi*[1]*, Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Speech, and Audio Processing Group, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{parvaneh.janbakhshi,ina.kodrasi,herve.bourlard}@idiap.ch

Abstract

Speech intelligibility is an important assessment criterion of the communicative performance of pathological speakers. To assist clinicians in their assessment, time- and cost-efficient automatic intelligibility measures offering a repeatable and reliable assessment are desired. In this paper, we propose to automatically assess pathological speech intelligibility based on a distance measure between the subspaces of spectral patterns of the pathological speech signal and of a fully intelligible (healthy) speech signal. To extract the subspace of spectral patterns we investigate two linear decomposition methods, i.e., Principal Component Analysis and Approximate Joint Diagonalization. Pathological speech intelligibility is then derived using a Grassman distance measure which quantifies the difference between the extracted subspaces of pathological and healthy speech. Experiments on an English database of Cerebral Palsy patients show that the proposed intelligibility measure is significantly correlated with subjective intelligibility ratings. In addition, comparisons to state-of-the-art measures show that the proposed subspace-based measure achieves a high performance with a significantly lower computational cost and without imposing any constraints on the speech material of the speakers.

Index Terms: spectral subspace, Grassman distance, Principal Component Analysis, Approximate Joint Diagonalization

1. Introduction

Speech is a very complex activity which can be significantly impaired due to pathologies caused by genetic influences, physical deformities, or neurological malfunctions. Many pathologies cause impairments in the speech production mechanism that result in reduced speech intelligibility and communicative ability [1]. As an index of pathology severity, functional limitation, and impairment progress, speech intelligibility assessment plays a crucial rule in clinical decision-making and monitoring.

Subjective listening tests are a gold standard for pathological speech intelligibility assessment. Such subjective assessments are not only time-consuming and costly, but they may also be influenced by the familiarity of the listener with the patient's speech pathology and the contextual/linguistic information available in the speech tasks under study [2]. To further assist clinicians in their assessments, automatic measures offering frequent, reliable, economical, and objective intelligibility assessment are required.

Automatic pathological speech intelligibility assessment approaches can be broadly categorized into i) blind approaches which do not require any healthy (intelligible) speech signals [3–10] and ii) non-blind approaches which exploit information about intelligible speech from healthy speakers [11–20].

Blind intelligibility assessment approaches typically refer to extracting several acoustic features that are believed to be correlated with intelligibility, such as the range of the fundamental frequency or the low-to-high modulation energy ratio [6, 7]. Intelligibility scores are then estimated by combining multiple features via feature selection and regression training [3–10].

Non-blind approaches encompass a wide range of approaches where healthy reference signals are exploited in different manners. For example, in [11] a single speaker-independent Gaussian Mixture Model (GMM) is trained on data of healthy speakers to create a healthy reference model. By adapting the parameters of this reference model, a GMM-based supervector is created to represent the pathological speech signal and the intelligibility score is obtained by training a regression model on the GMM-based supervector. A similar approach is followed in [12–14], with the difference consisting in using an iVector representation instead of a GMM-based supervector. In [15–19] healthy reference signals are needed to train an Automatic Speech Recognition (ASR) system. The ASR system is used to replace human listeners and pathological speech intelligibility is computed based on the word recognition rate. Finally, for the pathological speech intelligibility measure based on short-time objective intelligibility (P-ESTOI) in [20], healthy reference signals are used to create an intelligible utterance representation in the perceptually relevant octave band domain. After using Dynamic Time Warping (DTW) to align the pathological octave band representation to the intelligible representation, pathological speech intelligibility is directly computed as the divergence between the two aligned representations.

Many of the above-mentioned blind and non-blind approaches that rely on regression training have not followed a fair leave-one-subject-out evaluation paradigm, which might positively bias the reported results. Furthermore, the above-mentioned GMM-based, iVector-based, and ASR-based approaches require a large amount of healthy speech data. Additionally, ASR-based approaches are complex and might be unpredictable for severe patients [12]. Unlike these approaches, P-ESTOI does not suffer from such drawbacks since it does not require any regression training or a large amount of healthy speech data. However, P-ESTOI relies on time-alignment. Time-alignment using DTW might fail for severe patients, its computational cost is high when aligning long utterances, and it intrinsically requires healthy and pathological speakers uttering the same speech material, and hence, it cannot be used in phonetically unbalanced scenarios.

Motivated by the success of P-ESTOI, while aiming to avoid time-alignment and its inherent drawbacks, in this paper we propose an automatic pathological intelligibility measure exploiting spectral bases of the octave band representations of intelligible and pathological speech. We propose to find the subspaces of spectral patterns characterizing intelligible

http://dx.doi.org/10.21437/Interspeech.2019-2791

(healthy) and pathological speech using linear decomposition methods such as Principal Component Analysis (PCA) or Approximate Joint Diagonalization (AJD). The distance between the two subspaces is quantified using a Grassman distance measure and used to predict the intelligibility score of the pathological speaker. Experimental results on the Universal Access (UA) speech database [21] of Cerebral Palsy (CP) patients show that the proposed measure yields significant correlations with subjective intelligibility scores, while avoiding time-alignment, large amounts of training data, and being applicable to phonetically unbalanced scenarios.

2. Subspace-Based Pathological Speech Intelligibility Assessment

Speech spectrograms can be well approximated using low-rank matrices constructed based on low-dimensional spectral patterns. In the context of pathological speech intelligibility assessment, we hypothesize that the spectral patterns characterizing intelligible (healthy) speech differ from the spectral patterns characterizing pathological speech, with the difference increasing as pathological speech intelligibility decreases. To automatically assess speech intelligibility using spectral patterns, we propose to i) compute spectral bases characterizing an intelligible (healthy) utterance, ii) compute spectral bases characterizing the test (pathological) utterance, and iii) compute a distance measure between the healthy and pathological spectral bases. A schematic representation of the proposed pathological speech intelligibility measure is presented in Figure 1.

In the remainder of this section, the computation of the spectral bases and of the distance measure are presented. Furthermore, insights on the computational complexity reduction that is achieved using the proposed measure instead of P-ESTOI are also provided.

2.1. Computing intelligible spectral bases

As in P-ESTOI [20], in order to obtain a simplified internal representation resembling the transform properties of the auditory system, signals are first transformed to the time-frequency (TF) domain using one-third octave band analysis. Let $\mathbf{H}_s$ denote the $(J \times M_s)$-dimensional TF representation of an utterance from healthy speaker s, with J the number of octave bands and M_s the number of time frames. Searching for spectral patterns characterizing an intelligible (healthy) utterance, we propose to project each time frame in $\mathbf{H}_s$ into a set of J-dimensional spectral bases vectors $\mathbf{u}_k$, $k = 1, 2, \ldots, B$, with $B < J$. To obtain meaningful spectral bases vectors, multiple representations of utterances by different healthy speakers should be taken into account, such that the spectral bases can capture patterns

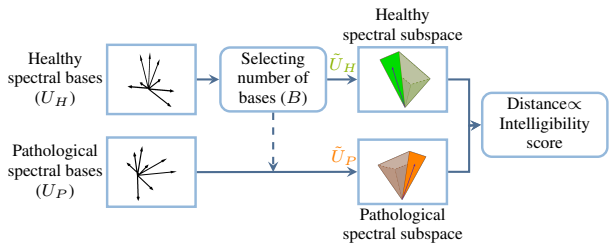

Figure 1: *Schematic representation of the proposed subspace-based pathological speech intelligibility measure.*

which are specific to intelligible speech but are independent of the particular speaker. A large number of matrix decomposition techniques exist to solve such a problem. As presented in Sections 2.1.1 and 2.1.2, we will use here the PCA and AJD techniques.

2.1.1. Principal Component Analysis

To take into account multiple healthy speakers, we consider the octave band representation of an utterance from all available healthy speakers and compute the average long-term spectral correlation $\mathbf{\Phi}_H$, i.e.,

$$\mathbf{\Phi}_H = \sum_{s=1}^{c} \frac{1}{M_s} \mathbf{H}_s \mathbf{H}_s^T, \tag{1}$$

with c being the number of available healthy speakers. The data are assumed to be mean-centered before computing the long-term spectral correlations. In order to obtain healthy spectral bases, the eigenvalue decomposition (EVD) of $\mathbf{\Phi}_H$ is first computed, i.e.,

$$\mathbf{\Phi}_H = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T, \tag{2}$$

with $\mathbf{U}$ being the $J \times J$-dimensional matrix of eigenvectors $\mathbf{u}_i$ and $\mathbf{\Sigma}$ being the $J \times J$-dimensional diagonal matrix of eigenvalues σ_i assumed to be sorted in descending order. The $(J \times B)$-dimensional matrix of spectral bases $\tilde{\mathbf{U}}_H$ is then defined as the first B eigenvectors in $\mathbf{U}$. Clearly, the number of considered spectral bases B is a hyper-parameter of the proposed technique. The choice of B is discussed in Section 2.1.3.

2.1.2. Approximate Joint Diagonalization

If spectral correlations from different healthy speakers differ significantly, computing spectral bases using the average of the long-term spectral correlations in (1) might yield spectral patterns that do not offer a reasonable approximation to the different representations. Hence, instead of averaging spectral correlations from all available healthy speakers, we also propose to compute the healthy spectral bases by means of AJD [22]. AJD computes a $J \times J$-dimensional orthonormal matrix $\mathbf{V}$ which aims to simultaneously transform all available spectral correlations $\frac{1}{M_s}\mathbf{H}_s\mathbf{H}_s^T$, $s = 1, 2, \ldots, c$, to the diagonal form based on a diagonality criterion [22], i.e.,

$$J(\mathbf{V}) = \sum_{s=1}^{c} \text{Off}\,(\mathbf{V}^T \frac{1}{M_s} \mathbf{H}_s \mathbf{H}_s^T \mathbf{V}), \tag{3}$$

where $\text{Off}\{\mathbf{A}\}$ denotes the sum of the squares of the off-diagonal elements of the $(N \times N)$-dimensional matrix $\mathbf{A}$, i.e.,

$$\text{Off}(\mathbf{A}) = \sum_{1 \leqslant i \neq j \leqslant N} |a_{ij}|^2, \tag{4}$$

with a_{ij} the $(i,j)^{th}$ entry of $\mathbf{A}$. After appropriately sorting the column vectors in $\mathbf{V}$, the $(J \times B)$-dimensional matrix of healthy spectral bases $\tilde{\mathbf{U}}_H$ is defined as the first B vectors in $\mathbf{V}$. For estimating $\mathbf{V}$, the Jacobi Angles algorithm [22] is used. Similarly to before, the number of considered spectral bases B is a hyper-parameter of the proposed technique.

2.1.3. Choosing the number of spectral bases B

The number of spectral bases B affects the performance of the proposed intelligibility measure. While on the one hand we would like to use a large number of spectral bases B to bet-

ter approximate the available spectral correlations, on the other hand, we would like to use a small number of spectral bases B to ensure that only spectral patterns important for speech intelligibility are being captured (rather than patterns describing extraneous variations such as speaker variability or noise). Hence, there is an inherent trade-off associated with selecting the number of bases B.

Dimensionality reduction techniques typically select B based on a user-defined threshold on the percentage of overall variance explained. Instead of selecting B based on the percentage of overall variance explained (which requires the threshold as an additional hyper-parameter to be optimized), we propose to automatically select B by adapting the L-curve method originally proposed for regularized least-squares techniques [23]. To apply the L-curve method, we compute the reconstruction error of the healthy spectral correlations for different bases number B. Due to the inherent trade-off between the reconstruction error and B, the plot of the reconstruction error versus B typically has an L-shape, with the corner of the L-curve (i.e., point of maximum curvature) representing a reasonable compromise between simultaneously minimizing the reconstruction error and keeping the number of spectral bases B as low as possible. The corner point is automatically determined using the triangle method [24].

When healthy spectral bases are derived using PCA, the reconstruction error can be straight-forwardly computed as

$$\epsilon_{\text{PCA}}(B) = \left\| \mathbf{\Phi}_H - \sum_{i=1}^{B} \sigma_i \mathbf{u}_i \mathbf{u}_i^T \right\|_F^2 = \sum_{i=B+1}^{J} \sigma_i, \quad (5)$$

with $\|\cdot\|_F$ denoting the matrix Frobenius norm. When healthy spectral bases are derived using AJD, we define the reconstruction error using the minimum least-squares distortion criteria [25], i.e.,

$$\epsilon_{\text{AJD}}(B) = \sum_{s=1}^{c} \min_{\mathbf{\Lambda}_s} \left\| \frac{1}{M_s} \mathbf{H}_s \mathbf{H}_s^T - \mathbf{V}_B \mathbf{\Lambda}_s \mathbf{V}_B^T \right\|_F^2, \quad (6)$$

where $\mathbf{V}_B$ denotes the matrix constructed from the first B columns of $\mathbf{V}$ found in (3) and $\mathbf{\Lambda}_s$ denotes a $(B \times B)$-dimensional diagonal matrix. Using some mathematical manipulations (not presented here due to space constraints) $\epsilon_{\text{AJD}}(B)$ can be computed as:

$$\epsilon_{\text{AJD}}(B) = \sum_{s=1}^{c} \text{Tr}\big((\frac{1}{M_s} \mathbf{H}_s \mathbf{H}_s^T)^2\big) - \sum_{i=1}^{B} {d_i}^2, \quad (7)$$

where d_i is the i^{th} diagonal element of $\sum_{s=1}^{c} \mathbf{V}^T \frac{1}{M_s} \mathbf{H}_s \mathbf{H}_s^T \mathbf{V}$.

2.2. Computing test spectral bases

To derive the pathological speech intelligibility measure, spectral bases of the test pathological representation need to be compared to the intelligible spectral bases computed in Section 2.1. Denoting by $\mathbf{P}_s$ the $(J \times M_s)$-dimensional TF representation of the test pathological utterance from speaker s, the $(J \times B)$-dimensional test pathological spectral bases $\tilde{\mathbf{U}}_P$ are computed based on the EVD of $\frac{1}{M_s} \mathbf{P}_s \mathbf{P}_s^T$ (similarly as in Section 2.1.1).

2.3. Computing a distance measure between spectral bases

To predict the pathological speech intelligibility, a distance measure between the subspaces spanned by the columns of $\tilde{\mathbf{U}}_H$ and $\tilde{\mathbf{U}}_P$ needs to be defined. The spanned subspaces can be viewed as points on the Grassman manifold. The Grassman manifold has a Riemannian structure that allows the computation of many different distance measures on the manifold. While other subspace distance measures can be used, in this work we use the f-norm Chordal distance defined as

$$d_{CF} = 2\sqrt{\sum_{i=1}^{B} sin^2(\theta_i/2)}, \quad (8)$$

where θ_i denotes the i^{th} principal angle between subspaces [26]. The final intelligibility score for each patient is obtained as the mean of the Chordal distance values across all considered utterances.

2.4. Complexity analysis

In this section, we provide some insights on the complexity reduction that is achieved when using the proposed subspace-based method instead of P-ESTOI.

The proposed subspace-based measure first needs to compute covariance matrices with a complexity of $\mathcal{O}(J^2 M)$, where M denotes the number of time-frames [27]. In addition, the complexity of the PCA decomposition based on the EVD is $\mathcal{O}(J^3)$ [28]. Similarly, the complexity of the AJD decomposition is also $\mathcal{O}(J^3)$ [29]. Hence, the proposed subspace-based measure (using either PCA or AJD) has a computational complexity of $\mathcal{O}(J^2 M + J^3)$.

When using P-ESTOI, the burden on the computational complexity arises due to using DTW. The DTW algorithm has a computational complexity of $\mathcal{O}(MN)$, with M and N being the number of time frames in the two octave band representations being aligned [30]. Additionally, for each iteration step of DTW, a frame-wise Euclidean distance with complexity $\mathcal{O}(J)$ needs to be computed. Hence, assuming $M = N$, the overall complexity of P-ESTOI is $\mathcal{O}(JM^2)$. Since $M \gg J$ (particularly for long utterances), using the proposed subspace-based measure instead of P-ESTOI reduces the computational complexity by a factor of M (i.e., from $\mathcal{O}(JM^2)$ to $\mathcal{O}(J^2 M + J^3)$), which can be advantageous when using such automatic measures for real-time feedback and assistance of clinicians.

3. Experimental Results

In this section, the performance of the proposed intelligibility measure is investigated and compared to state-of-the-art approaches.

3.1. Database

We consider the UA speech database with 763 isolated words from 15 English-speaking CP patients (11 males, 4 females) and 13 healthy speakers (9 males, 4 females) [21]. For each speaker, 155 isolated words were repeated 3 times, whereas the remaining words were uttered only once. The subjective intelligibility scores of patients range from 2% to 95%. Since multi-channel recordings are available, we consider the recordings of the 5th channel for our evaluation. Furthermore, an energy-based voice activity detection [31] is used to extract speech-only segments.

3.2. Evaluation measures, considered scenarios, and state-of-the-art approaches

To evaluate the performance, the automatically estimated intelligibility and the subjective intelligibility scores are compared in terms of the Pearson correlation coefficient (R) and

the Spearman rank correlation coefficient (R_s) along with their p-values (significance analysis).

To assess whether the length of the considered utterances has any effect on the computed intelligibility measures, the following scenarios are investigated:

i) Word-level analysis where intelligibility scores are computed for each word and the final intelligibility of the patient is obtained as the mean across all word-level intelligibility scores.

ii) Text-level analysis where intelligibility scores are computed for every 100 words concatenated to create a longer utterance. Concatenating words this way yields in total 8 utterances for each speaker. The final intelligibility of the patient is obtained as the mean across all text-level intelligibility scores.

For scenarios i) and ii) the utterances uttered by the patients and healthy speakers are the same. To assess whether phonetic variability between the utterances of patients and healthy speakers have any effect on the computed intelligibility measures, we additionally investigate the following scenarios:

iii) Text-level analysis with possibly common words where the 763 available words are randomly divided into two subsets of equal size. One subset is used for the healthy speakers whereas the other subset is used for the patients. Since some words are repeated in the database, the utterances of healthy speakers and patients overlap in terms of their phonetic content but are not the same. The random division of words into two subsets is repeated 10 times and the reported correlation is the average correlation across all repetitions whereas the reported p-value is the maximum value obtained from all repetitions.

iv) Text-level analysis without common words where a similar procedure as in iii) is followed, excluding however any common words between the healthy speakers and patients.

The proposed intelligibility measure (using either PCA or AJD) is compared to two state-of-the-art approaches, i.e., P-ESTOI [20] and iVector-based regression [14]. The algorithmic settings used for P-ESTOI are the same as in [20]. It should be noted that P-ESTOI cannot be used in scenarios iii) and iv) since it requires the phonetic content between the healthy speakers and patients to be exactly the same. For the iVector-based regression approach, we report the results from [14] where the authors have evaluated the approach on the same database with a leave-one-out strategy. It should be noted that the iVector-based regression approach has been implemented on the word-level scenario described in i).

Table 1: *Performance of the proposed and state-of-the-art measures on* 15 *English CP patients for different scenarios.*

Measures	R	p	R_S	p
i) Word-level				
P-ESTOI	0.94	2.5e−7	0.94	9.3e−7
iVector	0.74	–	–	–
AJD	−0.82	1.7e − 4	−0.88	1.8e−5
PCA	−0.83	1e−4	−0.88	1.8e−5
ii) Text-level				
P-ESTOI	0.93	9e−7	0.95	3e−7
AJD	−0.80	3.9e−4	−0.78	5.7e−4
PCA	−0.81	2.4e−4	−0.83	1.2e−4
iii) Text-level with common words				
AJD	−0.79	6e−4	−0.78	9.5e−4
PCA	−0.78	0.008	-0.76	0.019
iv) Text-level without common words				
AJD	−0.78	9.5e−4	−0.77	0.001
PCA	−0.7	0.009	−0.65	0.047

3.3. Results

Table 1 presents the correlation and p-values obtained by the proposed subspace-based measures (using either PCA or AJD), P-ESTOI, and the iVector-based approach for different scenarios. We consider the correlation values to be statistically significant if $p < 0.05$. It can be observed that the proposed subspace-based intelligibility measure using the PCA and AJD decompositions achieves *a high and significant correlation with the subjective intelligibility scores in all considered scenarios*. When the phonetic content between healthy speakers and patients is entirely different (i.e., scenario iv)), using the AJD-based decomposition appears to be more advantageous than using the PCA-based decomposition. Considering the word-level and text-level intelligibility assessment with balanced phonetic content, P-ESTOI gives the highest correlation values, which is to be expected since it takes both spectral and temporal distortions into account (while the proposed measure does not take any temporal distortion into account). However, P-ESTOI is computationally more expensive than the proposed measure and cannot be used in phonetically unbalanced scenarios. Furthermore, the iVector-based approach yields the lowest Pearson correlation in word-level intelligibility analysis (R_s and p values are not reported in [14]).

In summary, it can be said that the proposed subspace-based measure (using either PCA or AJD) yields *high and significant correlations with subjective scores and can also be applied to scenarios with phonetic variability* between the healthy and pathological speakers. In such scenarios, *using the AJD decomposition outperforms using the PCA decomposition*. Additional experimental results (not presented here due to space constraints) suggest that the proposed subspace-based measure achieves a high performance *independently of the language, of the speech pathology, or of the choice of healthy speakers to compute the intelligible subspace*.

4. Conclusion

To automatically assess pathological speech intelligibility, we have proposed a measure based on the subspaces spanned by the spectral bases of the octave band representation of healthy and pathological speech signals. Once intelligible (healthy) and pathological subspaces are estimated by applying linear decompositions on the speech data, pathological speech intelligibility is quantified based on a Grassman distance between the two subspaces. Experimental results on the UA speech database of CP patients show that the proposed measure can obtain high correlations with subjective intelligibility scores, while being more computationally efficient than state-of-the-art measures, not requiring a large amount of healthy data, and being applicable to phonetically unbalanced speech data between healthy and pathological speakers.

5. Acknowledgements

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 all project partners for a fruitful collaboration.

6. References

[1] P. Enderby, "Disorders of communication: Dysarthria," *Handbook of Clinical Neurology*, vol. 110, pp. 273–281, Jan. 2013.

[2] S. Landa, L. Pennington, N. Miller, S. Robson, V. Thompson, and N. Steen, "Association between objective measurement of the speech intelligibility of young people with dysarthria and listener ratings of ease of understanding," *International Journal of Speech-Language Pathology*, vol. 16, no. 4, pp. 408–416, Aug. 2014.

[3] J. C Kim, H. Rao, and M. A Clements, "Speech intelligibility estimation using multi-resolution spectral features for speakers undergoing cancer treatment," *Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 315–321, Oct. 2014.

[4] 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.

[5] 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.

[6] T. H. Falk, R. Hummel, and W. Y. Chan, "Quantifying perturbations in temporal dynamics for automated assessment of spastic dysarthric speech intelligibility," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May. 2011, pp. 4480–4483.

[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, Jun. 2012.

[8] C. Fang, H. Li, L. Ma, and M. Zhang, "Intelligibility evaluation of pathological speech through multigranularity feature extraction and optimization," *Computational and Mathematical Methods in Medicine*, vol. 2017, Jan. 2017.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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. 10:1–10:21, 2015.

[15] M. Schuster, E. Nöth, T. Haderlein, S. Steidl, A. Batliner, and F. Rosanowski, "Can you understand him? Let's look at his word accuracy-automatic evaluation of tracheoesophageal speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, USA, Mar. 2005.

[16] M. Schuster, T. Haderlein, E. Nöth, J. Lohscheller, U. Eysholdt, and F. Rosanowski, "Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating," *European Archives of Oto-Rhino-Laryngology*, vol. 263, no. 2, pp. 188–193, Feb. 2006.

[17] A. Maier, M. Schuster, A. Batliner, E. Nöth, and E. Nkenke, "Automatic scoring of the intelligibility in patients with cancer of the oral cavity," in *Proc. 8th Annual Conference of the International Speech Communication Association*, Antwerp, Belgium, Aug. 2007, pp. 1206–1209.

[18] M. Windrich, A. Maier, R. Kohler, E. Nöth, E. Nkenke, U. Eysholdt, and M. Schuster, "Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma," *Folia Phoniatrica et Logopaedica*, vol. 60, no. 3, pp. 151–156, Mar. 2008.

[19] A. Maier, T. Haderlein, F. Stelzle, E. Nöth, E. Nkenke, F. Rosanowski, A. Schützenberger, and M. Schuster, "Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, Dec. 2009.

[20] 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.

[21] 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.

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

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

[24] J. Castellanos, S. Gómez, and V. Guerra, "The triangle method for finding the corner of the l-curve," *Applied Numerical Mathematics*, vol. 43, pp. 359–373, Dec. 2002.

[25] M. Wax and J. Sheinvald, "A least-squares approach to joint diagonalization," *IEEE Signal Processing Letters*, vol. 4, no. 2, pp. 52–53, Feb. 1997.

[26] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: A unifying view on subspace-based learning," in *Proc. 25th International Conference on Machine Learning*, Helsinki, Finland, Jul. 2008, pp. 376–383.

[27] V. Kwatra and M. Han, "Fast covariance computation and dimensionality reduction for sub-window features in images," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 156–169.

[28] M. Tammen, I. Kodrasi, and S. Doclo, "Complexity reduction of eigenvalue decomposition-based diffuse power spectral density estimators using the power method," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, AB, Canada, Apr. 2018, pp. 451–455.

[29] V. Kuleshov, A. Chaganty, and P. Liang, "Tensor Factorization via Matrix Factorization," in *Proc. 18th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 38, San Diego, California, USA, May. 2015, pp. 507–516.

[30] M. Meinard, *Information Retrieval for Music and Motion*. Springer, Berlin, Heidelberg, 2007, ch. Dynamic Time Warping, pp. 69–84.

[31] B. Paul, "PRAAT, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9, pp. 341–345, Jan. 2002.