PATHOLOGICAL SPEECH INTELLIGIBILITY ASSESSMENT BASED ON THE SHORT-TIME OBJECTIVE INTELLIGIBILITY MEASURE

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

Impaired speech intelligibility in motor speech disorders arising due to neurological diseases negatively affects the communication ability and quality of life of patients. Reliable and cost-effective measures to automatically assess speech intelligibility are necessary for the management of such disorders. In this paper, we propose to automatically assess the intelligibility of pathological speech based on short-time objective intelligibility measures typically used in speech enhancement, which however require a reference signal that is time-aligned to the test signal. We propose a method to create an utterance-dependent reference signal of intelligible speech from multiple healthy speakers. In order to assess intelligibility, the pathological speech signal is aligned to the created reference signal using dynamic time warping and the divergence between the two signals is quantified using either the short-time or the spectral correlation. Experiments on databases of English and French patients suffering from Cerebral Palsy and Amyotrophic Lateral Sclerosis show that the proposed intelligibility measures can obtain a high correlation with subjective intelligibility ratings, outperforming several state-of-the-art pathological speech intelligibility measures.

Index Terms— STOI, ESTOI, DTW, pathological speech intelligibility

1. INTRODUCTION

Dysarthria of speech results from disturbances of the muscular control on the movement mechanism necessary for the execution of speech [1] and is a common symptom of several neurological diseases such as Cerebral Palsy (CP) or Amyotrophic Lateral Sclerosis (ALS). Dysarthria affects several components of the speech production mechanism such as respiration, phonation, resonance, articulation, and prosody, yielding an abnormal quality of speech as well as a reduced intelligibility and communicative ability [2]. Speech intelligibility is an important clinical and social aspect in the management of dysarthric speakers since it helps to characterize the severity of the speech disorder and functional communicative performance [3]. The gold standard pathological speech intelligibility measure is based on subjective listening tests evaluating the percentage of words correctly understood by human listeners. However, such a measure is labor-intensive, costly, and is also affected by the listener's familiarity with the patient's speech disorder and by the contextual/linguistic cues available in connected speech [4]. Therefore, time-and cost-efficient automatic intelligibility measures offering a repeatable and reliable assessment are desired. In the past decade several approaches to the automatic estimation of pathological speech intelligibility have been proposed, which can be broadly categorized into 3 categories, i.e., i) automatic speech recognition (ASR)-based approaches, ii) acoustic modeling-based approaches, and iii) feature-based approaches.

In ASR-based approaches, ASR systems are trained on large databases of healthy speech signals and are used to replace human listeners. Using a German ASR system, in [5–9] it is shown that the word recognition rate for pathological speech is correlated to the subjective intelligibility measure for patients suffering from different voice disorders, such as cancer of the oral cavity or head and neck cancer. In addition to word recognition rate, frame-level scores are used in [10], where an ASR system trained on Dutch speech data is used for forced alignment of the speech with the target word. The drawbacks of such ASR-based approaches for automatic pathological intelligibility estimation are their complexity, their need for a large amount of data, and unpredictability for severe patients [11]. Furthermore, for ASR systems in forced alignment mode, transcription or phonemicization of pathological speech is additionally needed.

In acoustic modeling-based approaches, intelligibility is estimated based on the vectors representing the utterance in the acoustic space, such as iVectors trained on an English database [11] or French database [12]. In [13], the pathological acoustic models are characterized by maximum a posteriori (MAP) adaptation of a gaussian mixture model (GMM) trained on German healthy speech [13]. The acoustic model parameters are then used for intelligibility estimation of patients with different voice disorders.

Finally, feature-based approaches typically refer to the blind assessment (not requiring any reference signal or other information such as phone boundaries) of speech intelligibility by extracting several acoustic features such as pitch range or voiced frames percentage. Using feature selection and regression training, an intelligibility measure is then derived [14–19]. In many of these approaches, rigorous validation strategies have not been followed. A fair leave-one-subject-out paradigm or a separate test and train set have not been reported for feature selection or regression training, which may lead to over-fitting and performance overestimation.

One of the objective intelligibility measures commonly used in speech enhancement is the short-time objective intelligibility (STOI) measure, which has been successful in estimating the intelligibility of time-frequency (TF) weighted noisy speech [20–22]. Additionally, to estimate the intelligibility of speech contaminated by temporally modulated noise the extended STOI (ESTOI) has been proposed [23].

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.

Motivated by the success of STOI and ESTOI, in this paper we propose to use similar measures for pathological speech intelligibility assessment. However, direct application of such enhancement objective measures in speech pathology is difficult, since they are based on comparing time-aligned noisy and reference (clean) signals. While the pathological speech signal can be viewed here as a noisy signal, the reference signal, i.e., the non-impacted and fully intelligible version of the patients' speech signal, is clearly not available. As a consequence, we propose to use dynamic time warping (DTW) [24] to create an utterance-dependent reference signal from multiple healthy speakers. Furthermore, DTW is used to align the pathological speech signal to the reference signal and intelligibility measures are computed based on the short-time or spectral cross-correlation of the aligned signals. The motivation behind the resulting new measures, referred to as P-STOI or P-ESTOI, is that they have a simple structure and they take speech perception and distortion into account. Experiments on two different databases of French-speaking ALS patients and English-speaking CP patients show that the proposed measures yield high correlations with subjective intelligibility scores, outperforming several state-of-the-art feature-based approaches.

2. OVERVIEW OF STOI AND ESTOI

The STOI and ESTOI measures as defined in [20,23] require a clean and a degraded speech signal, which are assumed to be time-aligned. To estimate speech intelligibility, first one-third octave band analysis is applied to the TF representation of both clean and degraded signals, yielding in total J one-third octave bands. We denote the $J \times T$-dimensional time-aligned one-third octave band representations of the clean and degraded signals as $\mathbf{H}$ and $\mathbf{P}$, respectively, with T being the total number of frames. TF-units are denoted by $H_j(i)$ and $P_j(i)$, with j denoting the octave band index and i denoting the frame index.

In STOI, the intermediate intelligibility measure for the j^{th} octave band, denoted by $d_j^S(t)$, is calculated from a region of I consecutive TF-units, with $i \in \{t, (t+1), ..., (t+I-1)\}$ for $t \leq T-I+1$. Based on the energy of consecutive frames in $\mathbf{H}$, local normalization and clipping are applied to the corresponding frames in $\mathbf{P}$. Local normalization ensures that the energy of I consecutive TF-units of both clean and degraded representations is equal, whereas clipping ensures that the signal-to-distortion ratio (SDR) is lower bounded. The intermediate intelligibility measure $d_j^S(t)$ is then computed as the linear correlation coefficient between consecutive TF-units of the clean and degraded representations. Assuming independent frequency band contributions to intelligibility, the final STOI intelligibility measure, denoted by d^S, is obtained as the average of the intermediate measure over all frequency bands and time frames, i.e.,

$$d^S = \frac{1}{(T-I+1)J} \sum_{j,t} d_j^S(t), \tag{1}$$

with

$$d_j^S(t) = \frac{\sum_{i=t}^{t+I-1} \left(H_j(i) - \overline{H_j(i)}\right)\left(P_j(i) - \overline{P_j(i)}\right)}{\sqrt{\sum_{i=t}^{t+I-1} \left(H_j(i) - \overline{H_j(i)}\right)^2 \sum_{i=t}^{t+I-1} \left(P_j(i) - \overline{P_j(i)}\right)^2}}, \tag{2}$$

where $\overline{H_j(i)} = \frac{1}{I} \sum_{i=t}^{t+I-1} H_j(i)$ and $\overline{P_j(i)}$ is similarly defined.

Differently from STOI, ESTOI does not assume mutual independence between contributions of frequency bands to intelligibility and computes a spectral correlation instead of a temporal one. To calculate ESTOI, all I consecutive TF-units in each band of the clean and degraded representations are mean and variance normalized. The normalized TF-units are denoted by $\tilde{H}_j(i)$ and $\tilde{P}_j(i)$, with $i \in \{t, (t+1), ..., (t+I-1)\}$. For each time frame, the linear correlation coefficient between J frequency bands is computed. An intermediate intelligibility measure, denoted by $d^E(t)$ for $t \leq T-I+1$, is then computed as the average of the spectral linear correlation coefficients across I consecutive time frames. Finally, the ESTOI intelligibility measure, denoted by d^E, is calculated as the average of the intermediate measure over all frames, i.e.,

$$d^E = \frac{1}{(T-I+1)} \sum_t d^E(t), \tag{3}$$

with

$$d^E(t) = \frac{1}{I} \sum_{i=t}^{t+I-1} \frac{\sum_{j=1}^{J} \left(\tilde{H}_j(i) - \overline{\tilde{H}_j(i)}\right)\left(\tilde{P}_j(i) - \overline{\tilde{P}_j(i)}\right)}{\sqrt{\sum_{j=1}^{J} \left(\tilde{H}_j(i) - \overline{\tilde{H}_j(i)}\right)^2 \sum_{j=1}^{J} \left(\tilde{P}_j(i) - \overline{\tilde{P}_j(i)}\right)^2}}, \tag{4}$$

where $\overline{\tilde{H}_j(i)} = \frac{1}{J} \sum_{j=1}^{J} \tilde{H}_j(i)$ and $\overline{\tilde{P}_j(i)}$ similarly defined.

3. PATHOLOGICAL INTELLIGIBILITY ASSESSMENT USING P-STOI AND P-ESTOI

Pathological speech intelligibility is a measure of the influence of the speech production deficit of a patient on a listener's perceptual understanding [25], with pathological and healthy speech being differently perceived. We hypothesize that quantifying the divergence of a perceptual acoustic representation of pathological speech from healthy (intelligible) speech yields a reliable pathological intelligibility assessment technique. In this paper, we propose to use the simple perceptual acoustic representation used in STOI and ESTOI, i.e., the previously described one-third octave band representation. By computing either the short-time correlation or the spectral correlation between the octave band representations of a reference and time-aligned test signal, two estimates of speech intelligibility can be derived. For pathological intelligibility assessment, the test signal is the pathological signal while a time-aligned (fully intelligible) reference signal is not available. In the following, the time alignment and the method proposed to create utterance-dependent reference signals are described.

3.1. Time alignment

Let $\mathbf{X}_s$ denote the $J \times M$-dimensional octave band representation of an utterance from speaker s, with M denoting the total number of time frames. In addition, let $\mathbf{x}_s(m)$ denote the representation at time frame m, with $m \in [1, \ \ldots, \ M]$. Similarly, let $\mathbf{X}_p$ and $\mathbf{x}_p(n)$ denote the one-third octave band representations of the same utterance from another speaker p, with $n \in [1, \ \ldots, \ N]$ and N being the total number of time frames in $\mathbf{X}_p$. The representations $\mathbf{X}_s$ and $\mathbf{X}_p$ are typically not aligned (due to different speakers and speaking rates) and are generally of different lengths, i.e., $M \neq N$. These two representations are aligned through DTW, using a simple Euclidian distance as the cost function [24]. DTW finds T-dimensional warping paths $\boldsymbol{\phi}_{s,p}$ and $\boldsymbol{\phi}_{p,s}$, with $T \geq \max[M, N]$, such that the warped representations $\mathbf{X}_s(\boldsymbol{\phi}_{s,p})$ and $\mathbf{X}_p(\boldsymbol{\phi}_{p,s})$ are point-to-point aligned sequences.

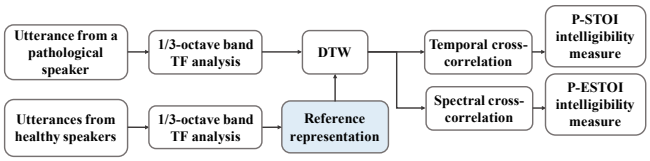

Fig. 1: *Block diagram of the proposed pathological intelligibility measures P-STOI and P-ESTOI. The reference representation (template) has been previously obtained from DTW-based clustering of healthy speaker templates (after 1/3-octave band TF analysis). A test (possibly pathological) utterance is then compared by DTW to the reference template to estimate temporal and spectral correlations, which are then used to calculate the P-STOI and P-ESTOI intelligibility measures (according to (1) and (3)).*

3.2. Utterance-dependent reference representations

For each considered utterance, a healthy speaker r is randomly selected, with $r \in \{1, \ldots, R\}$ and R being the total number of healthy speakers. Using DTW, the octave band utterance representation $\mathbf{X}_r$ is separately time-aligned with the representations from all remaining speakers. For each frame in $\mathbf{X}_r$, we extract all frames mapped to it by DTW from the representations of all remaining speakers. The representation for each reference frame is then created by taking the mean of all extracted aligned frames. The reference template for the considered utterance is then simply obtained by concatenating all reference frames so obtained.

It should be noted that using such an approach results in a reference representation that has the same length as the initial randomly selected representation $\mathbf{X}_r$. Our experimental results suggest that the computed P-STOI and P-ESTOI measures are not sensitive to the selected initial reference representation. In addition, our experimental results (not presented here due to space constraints) suggest that it is beneficial to use gender-specific reference representations, i.e., reference templates constructed using only healthy male (female) speakers when evaluating the intelligibility of male (female) patients. However, if the number of available healthy speakers is too small, it is more beneficial to use all speakers and create a single reference representation for both male and female patients.

3.3. Intelligibility assessment

To assess intelligibility, the one-third octave band representation for the considered test utterance is computed and aligned to the created reference template using DTW, with Euclidian distances as local scores. Due to different speaking rates, the aligned representations will obviously have repeated frames, i.e., after alignment, the shorter representation is likely to be expanded by repeating several frames.

In [26], it was shown that for diseases such as CP and ALS, the speaking rate did not show a high correlation with speech intelligibility. However, the repeated frames in the reference or patient representation will clearly affect the computed intelligibility measures. To discard the differences in speaking rates, these repeated frames are removed before computing the intelligibility measures. Denoting the TF-units of the aligned healthy reference and pathological test representations (with repeated frames discarded) as $H_j(t)$ and $P_j(t)$, P-STOI and P-ESTOI are computed using (1) and (3), respectively. P-STOI captures the impact of temporal distortions in the pathological speech on speech intelligibility whereas P-ESTOI focuses on the impact of spectral distortions.

The block diagram of the resulting pathological speech intelligibility measures P-STOI and P-ESTOI is illustrated in Figure 1.

4. EXPERIMENTAL RESULTS

In this section, the precision of the proposed P-STOI and P-ESTOI measures, as well as their generalisation properties across languages and neurological diseases are investigated here on French-speaking ALS patients and English-speaking CP patients. In addition, the proposed measures are compared to state-of-the-art feature-based measures proposed in [18]. The databases used are presented in Section 4.1, whereas the state-of-the-art feature-based measures and the evaluation criteria are presented in Section 4.2. Finally, the obtained results are presented and discussed in Section 4.3.

4.1. Databases

For evaluating the intelligibility of English-speaking CP patients we consider the Universal Access database [27]. Similarly to [18], 10 CP patients with spastic dysarthria (7 males, 3 females) are used for these experimental results, with subjective intelligibility scores ranging from 2% to 95%. For creating the reference representations, 13 healthy speakers (9 males, 4 females) are considered. Each patient and each healthy speaker read 763 isolated utterances, with a 7-channel microphone array used for the recordings. For the automatic intelligibility assessment, the recordings of the 5th channel have been considered. In order to extract speech-only segments, an energy-based voice activity detection (VAD) [28] has been used. Details on the database and the conducted subjective intelligibility test can be found in [27].

For evaluating the intelligibility of French-speaking ALS patients we consider recordings of 10 ALS patients (5 males, 5 females) from the University of Geneva and of 41 healthy speakers (19 males, 22 females) from the University of Paris III. The data has been recorded based on the MonPaGe speech screening protocol [29]. For the automatic intelligibility assessment, recordings of 49 utterances from all patients and healthy speakers have been considered. Manual VAD has been used to extract speech-only segments. For the subjective intelligibility assessment, the intelligibility module of the MonPaGe protocol is followed, where each patient is asked to utter 14 sentences containing 14 target words randomly selected from a list of 437 words. Six French native speakers with no prior experience with pathological speech were recruited to listen to the recorded sentences via headphones in a quiet environment and transcribe the target words. Listeners were allowed to listen to the sentences multiple times if desired. Subjective intelligibility scores were obtained based on the number of target words correctly understood by the listeners for each patient. The final subjective intelligibility score for each patient was computed as the average of the percentage of correctly understood target words across all listeners. The average intra-listener agreement for all words is 90.76%. The obtained subjective intelligibility scores of the patients range from 36% to 100%. It should be noted that only 2 patients have a subjective intelligibility score lower than 40%, with the remaining patients having a similar score that is higher than 90%.

4.2. State-of-the-art measures, evaluation, and settings

In order to compare P-STOI and P-ESTOI to the state-of-the-art measures, we consider several measures which have been shown to yield a high correlation with subjective intelligibility scores in [18], i.e., the kurtosis of the linear prediction residual $\mathcal{K}_{LP}$, the standard deviation of the zeroth order delta coefficient σ_{Δ}, the voicing percentage $\%\mathcal{V}$, the range of the fundamental frequency Δ_{f0}, and the low-to-high modulation energy ratio (LHMR). $\mathcal{K}_{LP}$ aims at characterizing vocal source excitation atypicality, σ_{Δ} aims at characteriz-

ing short-term temporal dynamics, $\%\mathcal{V}$ and Δ_{f0} aim at characterizing disordered prosody, and LHMR aims at characterizing long-term temporal dynamics. Before computing these measures, the VADs described in Section 4.1 have been used. The linear prediction residual and the voicing percentage have been computed using Praat [28], σ_{Δ} and Δ_{f0} have been computed using the Speech Signal Processing Python package [30], and LHMR has been computed using [31]. It should be noted that the used VAD and all implementation details for the different measures have not been reported in [18], hence, they might be different from the ones used in this paper.

To evaluate all considered measures, we use the Pearson correlation coefficient (R) and the Spearman rank correlation coefficient (R_s) between the automatically estimated intelligibility (as the mean across all considered utterances) and the subjective intelligibility scores. In addition, the p-values (significance analysis) for both correlation coefficients are also presented.

To compute P-STOI and P-ESTOI, the TF analysis is performed using a 32 ms Hamming window with an overlap of 50%. The number of consecutive frames I considered for the correlation is 15 and the number of one-third octave bands J is 15. Furthermore, in P-STOI, the SDR is lower-bounded by -15 dB.

Computing our intelligibility measures requires selecting an initial reference representation $\mathbf{X}_r$ (cf. Section 3), which might affect the final computed intelligibility scores. To analyze the sensitivity of the computed intelligibility measures, we repeat the process of creating a reference representation and of computing the final intelligibility scores using a different healthy speaker for the initial reference representation.

The presented correlation analysis for P-STOI and P-ESTOI are the mean and standard deviation of the correlations obtained for different reference representations. In addition, the presented p value is the maximum p obtained for different reference representations (i.e., the worst-case performance in terms of significance analysis).

4.3. Intelligibility assessment of dysarthric speech

Table 1 presents the correlation values R and R_s along with the corresponding p values for all considered measures and databases. The bold entries in the table indicate significant correlations, i.e., $p < 0.05$. Overall it can be observed that the proposed P-STOI and P-ESTOI measures achieve a *high and significant Pearson correlation* with the subjective intelligibility scores for both the CP ($R = 0.90$ and $R = 0.95$, $p < 0.05$) and ALS ($R = 0.87$ and $R = 0.95$, $p < 0.05$) databases. In addition, both measures achieve a *high and significant Spearman correlation* for the CP database ($R_S = 0.82$ and $R_S = 0.91$, $p < 0.05$), while the Spearman correlation coefficients for the ALS database *are not statistically significant* ($p > 0.05$). This suggests that there is not enough evidence to conclude that there is a monotonic relationship between the estimated and subjective intelligibility scores of the ALS patients. The low Spearman correlation for the ALS patients can be attributed to the highly skewed distribution of the subjective intelligibility scores for these patients, with 2 patients having similar subjective intelligibility scores lower than 40% and the remaining 8 patients having a similar subjective intelligibility score larger than 90%.

Table 1: *Performance of the proposed and state-of-the-art measures on the English CP and French ALS databases.*

Measures	R	p	R_S	p
English CP database				
P-STOI	**0.90 ± 0.004**	5e−4	**0.82 ± 0.002**	7e−3
P-ESTOI	**0.95 ± 0.004**	4.3e−5	**0.91 ± 0.000**	2e−4
$\mathcal{K}_{LP}$	0.41	0.23	0.42	0.23
σ_{Δ}	0.45	0.20	0.51	0.13
$\%\mathcal{V}$	-0.40	0.25	-0.58	0.08
Δ_{f0}	**-0.70**	0.02	-0.61	0.06
$LHMR$	-0.55	0.09	-0.54	0.10
French ALS database				
P-STOI	**0.87± 0.012**	2e−3	0.37 ± 0.004	0.33
P-ESTOI	**0.95± 0.006**	5.6e−5	0.43 ± 0.033	0.32
$\mathcal{K}_{LP}$	0.12	0.74	0.05	0.88
σ_{Δ}	**0.76**	0.01	0.48	0.16
$\%\mathcal{V}$	**-0.90**	2e−4	-0.60	0.06
Δ_{f0}	0.19	0.58	0.05	0.88
$LHMR$	**-0.69**	0.03	-0.46	0.18

Comparing the performance of P-STOI to P-ESTOI, it can be observed that P-ESTOI yields the best performance on both databases, suggesting that taking into account the different contribution of different frequency bands is beneficial for pathological speech intelligibility assessment.

Comparing the proposed measures to the considered state-of-the-art measures, it can be observed that using P-STOI and P-ESTOI yields *high correlations for both databases (i.e., for both considered languages and diseases)*, whereas the other considered measures either yield a significantly high correlation only for one of the considered databases (e.g., $\%\mathcal{V}$, Δ_{f0}, and σ_{Δ}) or yield a low correlation for both databases (e.g., $\mathcal{K}_{LP}$). The fundamental advantage of P-ESTOI and P-STOI over the other feature-based measures is that they rely on comparing a perceptual representation of the pathological speech to a reference perceptual representation of intelligible (healthy) speech, resulting in a high performance independently of the language or of the neurological disease.

Finally, it should be noted that the correlation values of the state-of-the-art measures on the CP database reported in Table 1 are not the same as the ones reported in [18]. This might be due to differences in the used VAD or in the implementation of the considered measures as described in Section 4.2. However, even when considering the correlation values reported in [18], P-ESTOI and P-STOI show improvements of up to 7%, 10% and 2%, 1% in Pearson and Spearman correlation over the best performing measure in [18]. It should however be mentioned that while the proposed P-STOI and P-ESTOI measures are reference-based measures, the considered state-of-the-art measures from [18] do not require a reference signal. Hence, in the future, we also aim to compare P-STOI and P-ESTOI to other reference-based measures.

5. CONCLUSION

Extending over the usual STOI and ESTOI intelligibility measures (where a reference signal is supposed to be available), we have proposed here the P-STOI and P-ESTOI measures to automatically assess pathological speech intelligibility. These measures rely on creating an utterance-dependent reference representation in one-third octave bands from healthy speakers. Intelligibility measures are then computed by quantifying the divergence of the pathological speech representation from the reference representation in terms of either the short-time or the spectral correlation. Experimental results on databases of English CP and French ALS patients have shown that P-STOI and P-ESTOI obtain a high correlation with subjective intelligibility scores, also yielding a higher correlation than several state-of-the-art measures. By relying on a reference representation created from multiple healthy speakers, P-STOI and P-ESTOI show a high performance, independently of the language or of the neurological disease.

6. REFERENCES

[1] J. R. Duffy, *Motor Speech Disorders: Clues to Neurologic Diagnosis, Diagnosis and Treatment Guidelines for the Practicing Physician*. Humana Press, Totowa, NJ, 2000, ch. Parkinson's Disease and Movement Disorders, pp. 35–53.

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

[3] K. C. Hustad, "Estimating the intelligibility of speakers with dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 58, no. 3, pp. 217–228, Feb. 2006.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] C. Middag, G. Van 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*, Baixas, France, Sep. 2008, pp. 1745–1748.

[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] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dallas, TX, USA, 2010, pp. 4214–4217.

[21] ——, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Feb. 2011.

[22] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Matching pursuit for channel selection in cochlear implants based on an intelligibility metric," in *Proc. 20th European Signal Processing Conference*, Bucharest, Romania, Aug. 2012, pp. 504–508.

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

[24] L. Rabiner and B. Juang, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993.

[25] B. M. Ansel and R. D. Kent, "Acoustic-phonetic contrasts and intelligibility in the dysarthria associated with mixed Cerebral Palsy," *Journal of Speech and Hearing Research*, vol. 35, no. 2, pp. 296–308, Apr. 1992.

[26] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 12, no. 2, pp. 246–269, Jun. 1969.

[27] 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.

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

[29] C. Fougeron, V. Delvaux, L. Ménard, and M. Laganaro, "The MonPaGe_HA database for the documentation of spoken French throughout adulthood," in *Proc. 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan, May 2018.

[30] P. Garner, "Speech signal processing (SSP) module," https://github.com/idiap/ssp, Jun. 2013.

[31] T. H. Falk, C. Zheng, and W. Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.