SYNTHETIC SPEECH REFERENCES FOR AUTOMATIC PATHOLOGICAL SPEECH INTELLIGIBILITY ASSESSMENT

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

Automatic pathological speech intelligibility measures are crucial to assist the clinical diagnosis and treatment of speech disorders. The recently proposed pathological short-time objective intelligibility (P-ESTOI) measure was shown to be very advantageous, yielding a high performance for several speech pathologies. However, to assess the intelligibility of an utterance from a patient, P-ESTOI relies on the availability of recordings of the same utterance by several healthy speakers such that an intelligible reference model can be created. Such recordings are not always easily available, limiting the practical applicability of P-ESTOI. To be able to use P-ESTOI in such scenarios, in this paper we propose to use synthetic speech generated by state-of-the-art high-quality text-to-speech systems to create an intelligible reference model. Experimental results on a database of Cerebral Palsy patients show that the performance of P-ESTOI using synthetic speech references is comparable to using natural speech references, making P-ESTOI a flexible measure which does not require healthy speech recordings and which outperforms state-of-the-art pathological speech intelligibility measures.

Index Terms— P-ESTOI, TTS, Cerebral Palsy

1. INTRODUCTION

Many pathological conditions disrupt the speech production mechanism, resulting in impairments that encapsulate altered speech production in different dimensions such as phonation, articulation, respiration, and prosody. These conditions include hearing loss [1], head and neck cancers [2], and neurological disorders, e.g., Parkinson's disease, Amyotrophic Lateral Sclerosis, or Cerebral Palsy (CP) [3]. Speech intelligibility is an important clinical and social aspect in the treatment of pathological speakers, since it helps to characterize the severity of the speech impairment and the functional communicative performance [4]. To assist clinicians with objective tools, there has been a growing interest in the research community to develop reliable automatic pathological intelligibility assessment measures.

While many approaches to automatic intelligibility assessment have been proposed, approaches which exploit healthy (i.e., perfectly intelligible) speech signals have shown very promising results. In these approaches, healthy speech recordings are exploited in different manners. In [5–7], healthy speech is used to train an Automatic Speech Recognition (ASR) system. The ASR system is used to replace human listeners, and pathological speech intelligibility is automatically computed based on the word recognition rate. In [8], 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 [9–11], with the difference consisting in using an iVector representation instead of a GMM-based supervector. All abovementioned approaches require a very large number of healthy speech recordings from many different speakers, which might not be easily available for under-resourced languages. In addition, most of these approaches use regression training on a large number of features, increasing as a result the risk of over-fitting. We have recently proposed a successful pathological intelligibility measure based on the short-time objective intelligibility (P-ESTOI) [12], which does not require any training or a very large amount of healthy speech recordings [12]. However, for assessing the intelligibility of a sample utterance from a patient in P-ESTOI, recordings of the same utterance from multiple healthy speakers are needed such that an utterance-dependent reference model can be created. 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. To be able to assess intelligibility also in such scenarios, we have recently proposed the subspace-based measure (SIM) [13]. Unlike P-ESTOI which exploits spectro-temporal cues for intelligibility assessment, SIM ignores temporal cues and relies on a comparison of spectral bases spanning the pathological speech signal to spectral bases spanning healthy speech. Although SIM can be used in phonetically unbalanced scenarios, its performance is inherently lower than P-ESTOI since temporal cues are not taken into account for intelligibility assessment [13].

In this paper, we propose to exploit synthetic speech generated by text-to-speech (TTS) systems to create intelligible reference models in P-ESTOI such that P-ESTOI becomes a flexible measure which can also be used in phonetically unbalanced scenarios (i.e., in scenarios where recordings from several healthy speakers uttering the same utterances as the pathological speaker are not available). This idea is motivated by the substantial progress made in the TTS field to generate high-quality synthesized speech capturing characteristics of intelligible natural speakers [14]. Using TTS systems as an "average" intelligible speaker has already been successfully exploited in the past for different applications. For example, in [15], synthetic speech is used for voice disorder detection by extracting acoustic features characterizing the deviation of the test speech signal from its synthesized counterpart. In [16], TTS systems are used to generate reference templates in template-based ASR systems,

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 Bastian Schnell for his assistance with the TTS system.

showing comparable ASR performance to generating reference templates using natural speech. To our knowledge, the suitability of synthetic speech references for pathological speech intelligibility assessment has never been investigated. Experimental results on a database of CP patients show that the performance of P-ESTOI using synthetic speech references is comparable to using natural speech references, making P-ESTOI a flexible measure which can also be used in phonetically unbalanced scenarios. In addition, it is shown that P-ESTOI using synthetic speech references outperforms SIM and other state-of-the-art measures.

2. OVERVIEW OF P-ESTOI

In this section, a brief overview of the state-of-the-art P-ESTOI measure is provided [12].

The computation of P-ESTOI relies on i) creating an utterance-dependent (intelligible) reference representation, ii) aligning the considered pathological representation to the reference representation using dynamic time warping (DTW), and iii) computing the spectral correlation between the two aligned representations to estimate the intelligibility. The method used to create the reference representation in the state-of-the-art P-ESTOI measure will be described in Section 3. The method we propose in this paper to create this reference representation will be described in Section 4. In the following, we assume that the reference representation is available and present the remaining steps required to compute P-ESTOI.

P-ESTOI operates in the perceptually relevant octave band domain. Hence, one-third octave band analysis is first applied to the time-frequency (TF) representation of all speech signals. We denote the $(J \times T)$–dimensional one-third octave band reference representation of a sample utterance n as $\mathbf{R}^n$, with J being the number of one-third octave bands and T being the number of time frames. Further, the one-third octave band representation of the same utterance from the pathological speaker k is denoted as $\mathbf{P}_k^n$. The representations $\mathbf{R}^n$ and $\mathbf{P}_k^n$ are unaligned and of different lengths. Using DTW, $\mathbf{P}_k^n$ is time-aligned to the reference representation $\mathbf{R}^n$. We denote TF-units of the aligned reference and pathological representations as $\mathrm{R}^n(j,i)$ and $\mathrm{P}_k^n(j,i)$, with j denoting the octave band index and i denoting the time frame index. To compute the pathological speech intelligibility, an intermediate intelligibility measure $d(t)$ is first computed from a region of I consecutive normalized TF-units, with $i \in \{t, (t+1), \ldots, (t+I-1)$ for $t \leq T-I+1$. Denoting by $\tilde{R}^n(j,i)$ and $\tilde{P}_k^n(j,i)$ the mean and variance normalized TF-units of each representation, $d(t)$ is computed as [12]

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

where $\overline{\tilde{R}^n(j,i)} = \frac{1}{J} \sum_{j=1}^{J} \tilde{R}^n(j,i)$ and $\overline{\tilde{P}_k^n(j,i)}$ is similarly defined. The intelligibility score of the sample utterance n from patient k (denoted as IS_k^n) is finally computed as

$$\mathrm{IS}_k^n = \frac{1}{(T-I+1)} \sum_t d(t). \tag{2}$$

Since P-ESTOI computes the spectral correlation of the time-aligned pathological and reference representations, it captures the impact that spectro-temporal distortions of the pathological utterance have on intelligibility.

3. HEALTHY SPEECH REFERENCES

As described in Section 2, P-ESTOI requires a reference representation $\mathbf{R}^n$ to estimate the pathological speech intelligibility for the sample utterance n. In the following, the computation of the reference representation $\mathbf{R}^n$ in the state-of-the-art P-ESTOI measure is described [12].

To construct $\mathbf{R}^n$ it is assumed that recordings of the same utterance by multiple healthy speakers are available. Let us denote the one-third octave band representation of utterance n from healthy speaker c by $\mathbf{H}_c^n$. For each utterance n, a healthy speaker r is randomly selected, with $r \in \{1, ..., R\}$ and R being the total number of healthy speakers. Using DTW, $\mathbf{H}_r^n$ is separately time-aligned with the representations of the same utterance from all the remaining healthy speakers. For each time frame in $\mathbf{H}_r^n$, all time frames mapped to it from the representations of all remaining healthy speakers are extracted and averaged. The reference representation $\mathbf{R}^n$ for utterance n is then simply obtained by concatenating all so-obtained averaged frames. Following such a procedure, the number of time frames in the reference representation $\mathbf{R}^n$ is dictated by the number of time frames in the representation of the utterance from the randomly selected reference speaker r.

4. SYNTHETIC SPEECH REFERENCES

As described in Section 3, to evaluate the intelligibility of an utterance from a pathological speaker, P-ESTOI creates a reference representation based on recordings of the same utterance from multiple healthy speakers. In practice however, such recordings are not always available. To make P-ESTOI a flexible measure which can be used in scenarios where such recordings are not available, in this section we propose to generate the reference representation using synthetic utterances generated with high-quality state-of-the-art TTS systems.

We propose to use a Deep Neural Network (DNN)-based TTS system inspired by the Merlin TTS system [17]. The Merlin TTS system has been used as a benchmark for assessing the quality of TTS systems in the *Blizzard Challenge* in 2016 [18] and 2017 [19]. It has been shown that such a system yields high-quality synthesized signals, outperforming systems based on Hidden Markov Models in terms of naturalness and intelligibility [17]. We train this system on multiple healthy speakers. For each sample utterance n in the pathological speech signal, we generate multiple synthesized reference utterances. The reference representation $\mathbf{R}^n$ is then computed following the same procedure as in Section 3. However, instead of using healthy speech recordings of the same utterance, we use synthesized speech of the same utterance from multiple TTS systems trained on multiple healthy speakers. Although following such an approach requires multiple healthy speech recordings to train appropriate TTS systems, it does not require healthy recordings of exactly the *same* utterances that are present in the pathological speech signal.

5. EXPERIMENTAL RESULTS

In this section, the performance of P-ESTOI using synthetic speech references as opposed to natural speech references is extensively investigated. In addition, the performance of P-ESTOI using synthetic speech references is compared to the performance of several state-of-the-art automatic pathological speech intelligibility measures.

5.1. Databases

The results presented in this section are based on the Universal Access database, which contains recordings of 15 US English-speaking CP patients (11 males, 4 females) [20]. In addition, we also use the recordings of 4 US English-speaking healthy speakers (2 males, 2 females) from this database. The subjective intelligibility scores of patients range from 2% to 95%. Each speaker utters 764 isolated words recorded by a 7-channel microphone array. For the automatic intelligibility assessment results presented in the following, the recordings of the 5th (arbitrary selected) channel have been considered. The sampling frequency of the recordings is 16 kHz. An energy-based voice activity detection is used to extract the speech-only segments [21].

For training the TTS systems, we consider the CMU ARCTIC database consisting of recordings of 1132 phonetically balanced utterances from 4 US English-speaking healthy speakers (2 males, 2 females) [22].

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

To compute reference representations from healthy speech signals, we use the 4 healthy speakers from the Universal Access database, i.e., $R = 4$ (cf. Section 3). To compute reference representations from synthetic speech signals, we train 4 TTS systems using the healthy recordings of the 4 healthy speakers in the CMU ARCTIC database. To this end, we use a DNN-based state-of-the-art Merlin TTS system in conjunction with the Festival front-end, two Bidirectional Long Short-Term Memory networks as duration and acoustic models, and the WORLD vocoder. For details on the TTS systems and the training procedure, the reader is referred to [23, 24]. By training a TTS system for each speaker, we get 4 speaker-dependent TTS systems. The remainder of the algorithmic settings used for the implementation of P-ESTOI are the same as in [12].

To evaluate the intelligibility assessment performance, we consider the Pearson correlation coefficient (R) and the Spearman rank correlation coefficient (R_s) along with their p-values (significance analysis) between the automatically estimated intelligibility and the subjective intelligibility scores. As previously mentioned, the computation of a reference representation in P-ESTOI (independently of whether natural or synthetic speech is used) requires selecting a random initial intelligible representation from the given set of natural or synthetic utterances. Hence, we repeat the computation of P-ESTOI multiple times using a different selection of the initial representation for creating the reference representation. The presented correlation values for P-ESTOI are the mean and standard deviation of the correlation values obtained for these different repetitions. The presented p-values are the maximum p-values obtained across all repetitions. To extensively analyze the proposed method and demonstrate its applicability, the following two scenarios are considered.

Phonetically balanced scenarios. In this scenario, we assume that all speakers (healthy and pathological) utter exactly the same utterances. All 764 utterances of the database are considered, and the final intelligibility score is computed as the mean across all utterance-level intelligibility scores. Only in such scenarios can the performance of P-ESTOI using synthetic speech references be compared to the performance of P-ESTOI using healthy speech references (since otherwise healthy speech reference models cannot be generated). The effect of the number of TTS systems used to generate reference representations is analyzed. In addition, the performance of P-ESTOI using synthetic references in this phonectically balanced scenario is compared to state-of-the-art measures, i.e., SIM [13], the ASR-based measure [11], and the iVector-based measure [11]. SIM is implemented based on Principal Component Analysis using the same settings as in [13]. For the ASR-based and iVector-based approaches we report the results from [11], where these approaches are evaluated on the same database of CP patients using a leave-one-out validation strategy.

Phonetically unbalanced scenarios. In this scenario, we assume that all speakers (healthy and pathological) utter different utterances. P-ESTOI using healthy speech references cannot be used in such scenarios since healthy reference models cannot be generated. Instead, the performance of P-ESTOI using synthetic speech references is compared to the performance of SIM, which is applicable to such phonetically unbalanced scenarios. The effect of the number of utterances that is available to estimate intelligibility is investigated.

5.3. Phonetically balanced scenarios

P-ESTOI using synthetic and natural references. In the following, the performance of P-ESTOI using synthetic speech references is compared to using healthy speech references. To analyze whether the performance of P-ESTOI is dependent on the number of TTS systems used to generate the reference representation, we investigate the performance when using 1, 2, 3, and 4 such TTS systems. Accordingly, this is compared to the performance of P-ESTOI using natural speech references generated from 1, 2, 3, and 4 healthy speakers. Since there are multiple ways of selecting, 1, 2, or 3 TTS systems or healthy speakers out of the available 4 TTS systems or healthy speakers, we have repeated the computation of P-ESTOI for each of these possible selections.

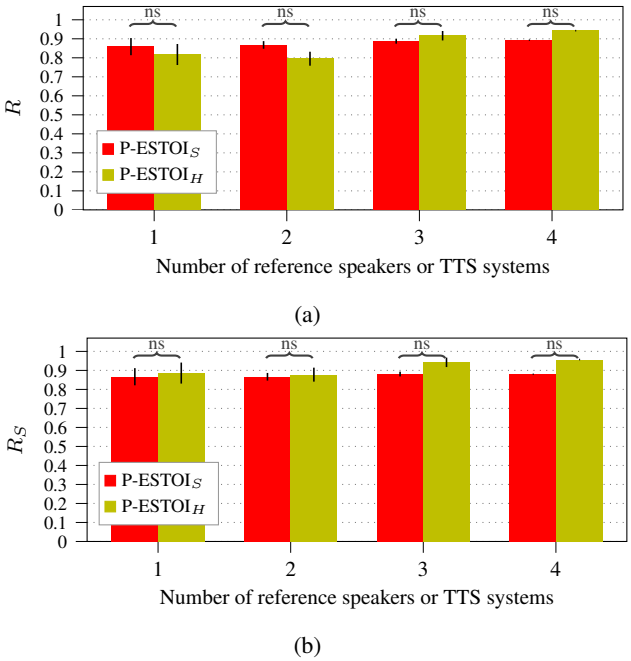

Fig. 1: *a) Pearson correlation R and b) Spearman rank correlation R_S using P-ESTOI$_S$ and P-ESTOI$_H$ for different number of TTS systems and healthy speakers. The columns and bars depict the mean and standard deviation of the correlation values across different selections of the set of TTS systems or healthy speakers. (ns) denotes non-significant differences between the correlation values of P-ESTOI$_S$ and P-ESTOI$_H$.*

Fig. 1 presents the Pearson and Spearman rank correlation between the subjective intelligibility scores and the P-ESTOI intelligibility measure using synthetic references (denoted as P-ESTOI$_S$) and healthy references (denoted as P-ESTOI$_H$) for different numbers of TTS systems or healthy speakers. The columns and bars in Fig. 1 present the mean and standard deviation of the correlation values across all repetitions. Although not presented in this figure, the correlation values obtained using both P-ESTOI$_S$ and P-ESTOI$_H$ across all repetitions are statistically significant (i.e., $p < 0.01$). It can be observed that the Pearson and Spearman correlation values obtained using P-ESTOI$_S$ and P-ESTOI$_H$ are both high and very similar, independently of the number of TTS systems or healthy speakers used to generate the reference representations. When using 1 or 2 TTS systems or healthy speakers, the Pearson correlation obtained with P-ESTOI$_S$ is slightly higher than the Pearson correlation obtained with P-ESTOI$_H$. In the remainder of the considered scenarios, the correlation values obtained with P-ESTOI$_S$ are slightly lower than the correlation values obtained with P-ESTOI$_H$.

To analyze whether the differences in the presented correlation values of P-ESTOI$_S$ and P-ESTOI$_H$ for each considered number of TTS systems or healthy speakers are statistically significant, we conduct a two-tailed dependent Steiger's Z-Test on all possible pairs of correlation values obtained across all repetitions [25]. The difference between the correlation values is considered significant when the obtained p-value is $p < 0.01$ in the majority (i.e., more than 50%) of the considered correlation pairs, otherwise this difference is considered to be non-significant (depicted by ns in Fig. 1). As shown in Fig. 1, there is no significant difference between P-ESTOI$_S$ and P-ESTOI$_H$ independently of the number of TTS systems or healthy speakers.

In summary, it can be said that the performance of P-ESTOI using synthetic speech references is very high and very similar to using natural speech references.

P-ESTOI and state-of-the-art measures. In the following, the performance of P-ESTOI using synthetic speech references generated from 4 TTS systems is compared to state-of-the-art pathological speech intelligibility measures, i.e., SIM, ASR-based, and iVector-based approaches. Since SIM is negatively correlated with intelligibility (cf. [13]), we present the absolute value of the correlation coefficients for this measure. Table 1 illustrates the performance of P-ESTOI$_S$, SIM, iVector-based, and ASR-based approaches in the considered phonetically balanced scenario. The Spearman rank correlation and p-values for the ASR-based and iVector-based approaches in [11] have not been reported. It can be observed that the proposed P-ESTOI$_S$ measure yields high and significant correlations (i.e., $p < 0.01$) with the subjective intelligibility scores, significantly outperforming the considered state-of-the-art measures.

5.4. Phonetically unbalanced scenarios

In this section, the performance of P-ESTOI using synthetic speech references is compared to the performance of SIM in phonetically unbalanced scenarios. To investigate the effect of the number of available utterances in estimating intelligibility, a set of utterances

Table 1: *Performance of P-ESTOI$_S$ and state-of-the-art measures in the phonetically balanced scenario.*

Measures	R	p	R_S	p
P-ESTOI$_S$	0.89 ± 0.008	1e−5	0.88 ± 0.020	6e−5
SIM	0.77	9e−4	0.84	7e−5
iVector	0.74	–	–	–
ASR	0.55	–	–	–

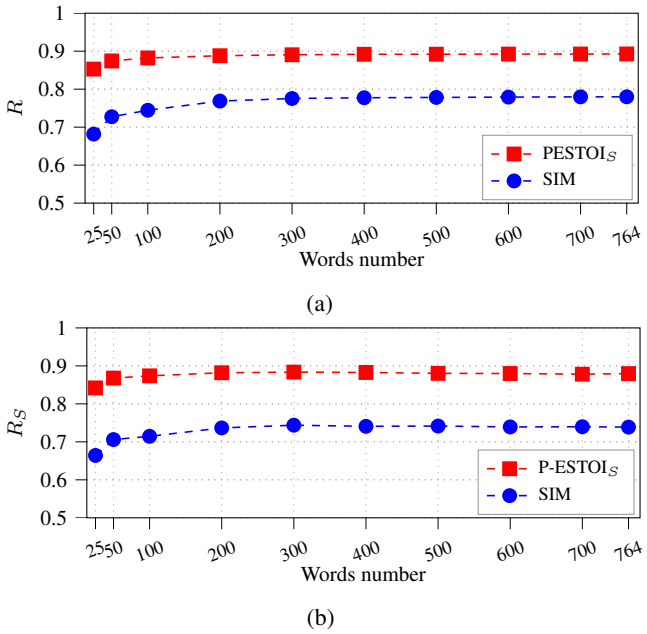

Fig. 2: *a) Pearson correlation R and b) Spearman rank correlation R_S using P-ESTOI$_S$ and SIM in phonetically unbalanced scenarios for different number of considered utterances.*

is randomly selected from the 764 available utterances for each speaker. The number of considered utterances ranges from 25 to 764. Clearly, there might be common utterances among speakers, however, these utterances are not exactly the same. The random selection of the set of utterances is repeated 100 times, and the final intelligibility score is obtained by averaging across all repetitions.

Fig. 2 presents the Pearson and Spearman rank correlation obtained using P-ESTOI$_S$ and SIM in the phonetically unbalanced scenario for different numbers of utterances. Although not presented in this figure, the correlation values obtained using P-ESTOI$_S$ are always statistically significant for each considered number of utterances (i.e., $p < 0.01$), whereas this is not always the case for SIM. It can be observed that independently of the considered number of utterances for intelligibility assessment, the correlation values obtained using P-ESTOI$_S$ are always higher than the ones obtained using SIM, demonstrating the advantages of the proposed method in phonetically unbalanced scenarios. Further, it can be observed that the correlation values obtained using P-ESTOI$_S$ quickly converge, showing that a relatively small number of utterances is necessary for P-ESTOI$_S$ to obtain a robust intelligibility assessment.

6. CONCLUSION

In this paper, we have proposed to create the reference representations required in P-ESTOI using high-quality synthetic utterances generated by state-of-the-art TTS systems. This way, P-ESTOI can be used in scenarios where healthy recordings of the same utterances as in the pathological speech signal are not available. Extensive experimental results on a database of CP patients have shown that the performance of P-ESTOI using synthetic speech references is comparable to using natural speech references. In addition, it has been shown that P-ESTOI using synthetic speech references significantly outperforms state-of-the-art automatic intelligibility measures, making P-ESTOI an advantageous and flexible measure which can be successfully used in a wide range of scenarios.

7. REFERENCES

[1] P. J. Flipsen and R. G. Parker, "Phonological patterns in the conversational speech of children with cochlear implants," *Journal of Communication Disorders*, vol. 41, no. 4, pp. 337–357, Jul. 2008.

[2] K. Mady, R. Sader, P. Hoole, A. Zimmermann, and H. Horch, "Speech evaluation and swallowing ability after intraoral cancer," *Clinical Linguistics and Phonetics*, vol. 17, no. 4–5, pp. 411–420, Jul. 2009.

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

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

[5] 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, May 2009.

[6] 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*, Brisbane, Australia, Sep. 2008, pp. 1745–1748.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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, May 2015.

[12] 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.

[13] ——, "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.

[14] F. Hinterleitner, C. Norrenbrock, and S. Möller, "Is intelligibility still the main problem? A review of perceptual quality dimensions of synthetic speech," in *Proc. 8th International Speech Communication Association Speech Synthesis Workshop*, Barcelona, Spain, Sep. 2013, pp. 147–151.

[15] G. K. Anumanchipalli, H. Meinedo, M. Bugalho, I. Trancoso, L. C. Oliveira, and A. W. Black, "Text-dependent pathological voice detection," in *Proc. 13th Annual Conference of the International Speech Communication Association*, Portland, USA, Sep. 2012, pp. 530–533.

[16] S. Soldo, M. Magimai-Doss, and H. Bourlard, "Synthetic references for template-based ASR using posterior features," in *Proc. 13th Annual Conference of the International Speech Communication Association*, Portland, USA, Sep. 2012, pp. 52–57.

[17] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *Proc. 9th International Speech Communication Association Speech Synthesis Workshop*, Sunnyvale, USA, Sep. 2016, pp. 202–207.

[18] S. King and V. Karaiskos, "The Blizzard challenge 2016," in *Proc. Blizzard Challenge Workshop*, Cupertino, USA, Sep. 2016.

[19] S. King, L. Wihlborg, and W. Guo, "The Blizzard challenge 2017," in *Proc. Blizzard Challenge Workshop*, Stockholm, Sweden, Aug. 2017.

[20] 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.

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

[22] J. Kominek and A. Black, "The CMU Arctic speech databases," in *Proc. 5th International Speech Communication Association Speech Synthesis Workshop*, Pittsburgh, USA, Jan. 2004, pp. 223–224.

[23] B. Schnell and P. N. Garner, "A neural model to predict parameters for a generalized command response model of intonation," in *Proc. 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, Sep. 2018, pp. 3147–3151.

[24] F. Marelli, B. Schnell, H. Bourlard, T. Dutoit, and P. N. Garner, "An end-to-end network to synthesize intonation using a generalized command response model," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brighton, UK, May 2019, pp. 7040–7044.

[25] J. H. Steiger, "Tests for comparing elements of a correlation matrix," *Psychological Bulletin*, vol. 87, no. 2, pp. 245–251, Mar. 1980.