Automatic Discrimination of Apraxia of Speech and Dysarthria using a Minimalistic Set of Handcrafted Features

Ina Kodrasi[*], *Michaela Pernon*[†], *Marina Laganaro*[†], *Hervé Bourlard*[*]

[*]Speech and Audio Processing Group, Idiap Research Institute, Martigny, Switzerland
[†]Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland
ina.kodrasi@idiap.ch

Abstract

To assist clinicians in the differential diagnosis and treatment of motor speech disorders, it is imperative to establish objective tools which can reliably characterize different subtypes of disorders such as apraxia of speech (AoS) and dysarthria. Objective tools in the context of speech disorders typically rely on thousands of acoustic features, which raises the risk of difficulties in the interpretation of the underlying mechanisms, over-adaptation to training data, and weak generalization capabilities to test data. Seeking to use a small number of acoustic features and motivated by the clinical-perceptual signs used for the differential diagnosis of AoS and dysarthria, we propose to characterize differences between AoS and dysarthria using only six handcrafted acoustic features, with three features reflecting segmental distortions, two features reflecting loudness and hypernasality, and one feature reflecting syllabification. These three different sets of features are used to separately train three classifiers. At test time, the decisions of the three classifiers are combined through a simple majority voting scheme. Preliminary results show that the proposed approach achieves a discrimination accuracy of 90%, outperforming using state-of-the-art features such as openSMILE which yield a discrimination accuracy of 65%.

Index Terms: formant frequencies, voiced segment duration, loudness, long-term average speech spectrum, temporal sparsity, SVM

1. Introduction

Different types of motor speech disorders (MSDs) are observed following various conditions of brain damage [1]. Apraxia of speech (AoS) and dysarthria refer to two distinct types of MSDs, presenting some specific but also overlapping clinical signs. On the one hand, AoS is considered to be a dysfunction of motor planning and is characterized by the presence of segmental distortions such as vowel distortion, inappropriate vowel lengthening, or reduced co-articulation, as well as several supra-segmental specificities leading to syllabification and an overall slow speech rate [1,2]. On the other hand, dysarthrias, although considered to encompass several disorders of motor speech execution (rather than planning), share similar clinical-perceptual signs with AoS such as segmental distortions and an overall slow speech rate [1–3]. The clinical differential diagnosis of AoS is usually based on the co-occurrence of signs that are not typically associated with dysarthrias, such as inconsistency of segmental distortions, groping, initiation problems, and syllabification [1,2,4]. Similarly, the clinical differential diagnosis of dysarthrias is based on clinical signs that are not typically associated with AoS, such as reduced loudness variation or hypernasality [2]. However, due to the overlapping of such signs and the difficulty of detecting them by ear, differential diagnosis is very hard to be achieved by non-experts and even expert inter-rater agreement is low [5–7]. To assist clinicians in the differential diagnosis and treatment of MSDs, it is therefore imperative to establish automatic tools which can reliably discriminate between AoS and dysarthria.

In the past decade, there has been a growing interest in the research community to develop objective tools to characterize speech disorders. To our knowledge, the majority of contributions deal with discriminating between healthy and impaired speech, with impairments arising due to dysarthrias or laryngeal pathologies. Several successful contributions for the automatic discrimination of healthy and dysarthric speech have been made, with a vast number of acoustic features being typically used such as the fundamental frequency, formant frequencies, jitter, shimmer, harmonics-to-noise ratio, or mel frequency cepstral coefficients (MFCCs) [8–11]. Discriminating between different types of speech impairments has not been as comprehensively studied, with seldom contributions dealing with discriminating between different types of laryngeal pathologies. Using many frame-level acoustic features, discrimination between two, three, and five laryngeal pathologies is done in [12–14] by means of Gaussian Mixture Models, Support Vector Machines (SVMs), or Hidden Markov Models.

The objective of this paper is to propose an automatic tool that can be used to discriminate between AoS and dysarthria. As previously mentioned, the recent use of machine learning approaches in the domain of speech disorders has led to a vast number of acoustic features (in the order of thousands) being typically employed [8, 15–17]. While this up-scaling of the number of features allows to capture many acoustic characteristics, it comes at the cost of serious difficulties in the interpretation of the underlying mechanisms. Further, using a large number of acoustic features poses the risk of over-adaptation to training data and weak generalization capabilities to test data. To overcome such limitations and motivated by the clinical-perceptual signs used to diagnose AoS and dysarthria, in this paper we propose to automatically characterize differences between AoS and dysarthria using only six handcrafted acoustic features. To capture segmental distortion differences between AoS and dysarthria, we propose to use the first formant frequency, the second formant frequency, and the duration of continuous voiced regions. To capture loudness and hyper-nasality differences, we propose to use the number of loudness peaks per second and the long-term average speech spectrum (LTAS). Finally, to capture syllabification differences, we propose to use the temporal sparsity of speech spectral coefficients. We train an SVM for each of these three feature sets, and as result, obtain three classification decisions per speaker. At test time, the decisions of the different SVMs are combined through a simple majority voting (MV) scheme.

Experimental results on a French database of patients suffering from AoS and dysarthria show that the proposed MV approach yields a high accuracy of 90%, outperforming the accuracy of the individual SVMs trained on the different feature sets. Further, these results show that the proposed approach significantly outperforms using an SVM with state-of-the-art feature sets such as openSMILE [18].

2. Automatic AoS and Dysarthria Discrimination

Fig. 1 depicts a schematic representation of the proposed approach for the automatic discrimination of AoS and dysarthria. As shown in this figure, we extract three sets of acoustic features, with the first one characterizing segmental distortions, the second one characterizing loudness and hyper-nasality, and the third one characterizing syllabification. Using these feature sets, three different SVMs are trained. At test time, the decisions of the three SVMs are combined through an MV scheme. In the remainder of this section, details on the proposed approach are provided.

2.1. Segmental distortion

The feature vector for SVM_1 is constructed by concatenating the statistics of formant frequencies and duration of continuous voiced regions.

Statistics of formant frequencies. As previously mentioned, vowel distortion and reduced co-articulation are some of the clinical signs used to diagnose AoS. The perception of the quality of vowels and co-articulation patterns is determined by several acoustic properties, with the variation of the first and second formant frequencies F_1 and F_2 being particularly important [19–21]. Hence, to characterize vowel distortions and co-articulation patterns, we propose to compute F_1 and F_2 based on linear predictive coding as in [18]. The mean, standard deviation, kurtosis, and skewness of F_1 and F_2 across time are used for the feature vector of SVM_1.

Statistics of the duration of continuous voiced regions. As previously mentioned, abnormalities in vowel duration can be observed in AoS. To capture such abnormalities, we propose to compute the duration of each continuous voiced region (i.e., each continuous region where the estimated fundamental frequency is greater than 0) as in [18]. The mean and standard deviation of the duration of all such regions found in the utterance are appended to the feature vector of SVM_1.

Hence, the final feature vector for SVM_1 is the 10-dimensional vector constructed by concatenating the statistics of formant frequencies and duration of continuous voiced regions.

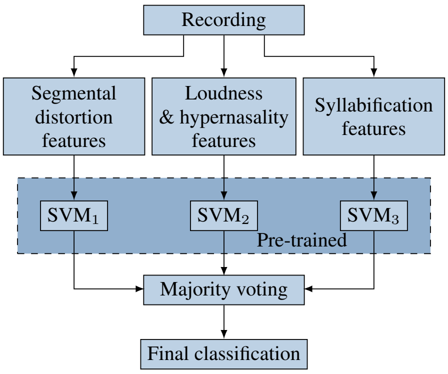

Figure 1: *Schematic representation of the proposed approach for the automatic discrimination of AoS and dysarthria.*

2.2. Loudness and hyper-nasality

The feature vector for SVM_2 is constructed by concatenating the number of loudness peaks per second and the LTAS.

Number of loudness peaks per second. As described in Section 1, reduced loudness variation is a clinical sign used to diagnose dysarthria. To characterize loudness variation, we propose to compute the number of loudness peaks per second. First, the auditory spectrum is computed as in [22] using the power of the Mel-band spectrum weighted with an equal loudness curve. By computing the loudness through summation over all auditory bands, the number of loudness peaks per second is extracted and used for the feature vector of SVM_2.

Long term average speech spectrum. As described in Section 1, hyper-nasality is another clinical sign used to diagnose dysarthria. Hyper-nasality may manifest itself as an atypical distribution of energy across the speech spectrum [23, 24]. To capture this cue, we propose to use the LTAS computed from the octave band representation. We use nine octave bands and the average speech power across time in each band is appended to the feature vector for SVM_2.

Hence, the final feature vector for SVM_2 is the 10-dimensional vector constructed by concatenating the number of loudness peaks per second and the LTAS.

2.3. Syllabification

The feature vector for SVM_3 is constructed by concatenating the statistics of temporal sparsity of the spectral coefficients.

Statistics of temporal sparsity. As previously mentioned, syllabification can be observed in AoS. Syllabification manifests itself as excessive pauses between phoneme transitions and words. Such excessive pauses imply that there is a consistent lack of energy in the speech spectral coefficients, which can be captured by means of the temporal sparsity [25].[1] As in [27], temporal sparsity is computed for each frequency through a maximum likelihood estimate of the shape parameter of a Chi distribution modeling the speech spectral magnitudes. A lower shape parameter describes more sparse signals, i.e., signals with excessive pauses due to syllabification. The mean, standard deviation, kurtosis, and skewness of the so-computed shape parameter across frequency are used to create the 4-dimensional feature vector for SVM_3.

2.4. Classification

Depending on the patient under consideration and the available speech material, the discriminative power of different features, and hence, the classification decision of the three classifiers, can vary. After training the SVMs, the decisions of the three classifiers are combined at test time through a simple MV scheme. As will be shown in Section 3, the performance of this combination scheme is higher than the performance of the individual SVMs.

[1] It should be noted that identifying a single perceptual characteristic that temporal sparsity reflects is not straight-forward. Apart from syllabification, temporal sparsity can reflect articulation deficiencies and inconsistent formant transitions [26, 27]. We have chosen to associate temporal sparsity with syllabification in this paper since syllabification is an established perceptual characteristic to identify AoS.

3. Results and Discussion

In this section, the performance of the proposed approach is investigated and compared to using an SVM with the state-of-the-art openSMILE features [18].

3.1. Database and preprocessing

The results presented in the following are based on a database collected at Geneva University Hospitals and University of Geneva. We consider 20 patients, with 10 patients (4 male, 6 female) diagnosed with AoS and the remaining 10 patients (4 male, 6 female) diagnosed with dysarthria. All AoS patients have suffered a stroke, 7 of the patients with dysarthria suffer from Parkinson's disease, and the remaining 3 patients suffer from Amyotrophic Lateral Sclerosis. The age of the AoS patients ranges from 24 to 72 years old, with a mean age of 53 years old. The age of the dysarthria patients ranges from 55 to 83 years old, with a mean age of 73 years old. Since there is an age mismatch between the two groups of patients, a regression approach is considered to validate the classification results (cf. Section 3.4).

For each patient, the overall severity of the speech impairment was evaluated by a speech pathologist using the perceptive score of BECD[2] [28]. The perceptive BECD score ranges from 0 (no impairment) to 20 (severe impairment), reflecting impairments in different dimensions such as voice quality, phonetic production, prosody, or intelligibility [28]. The perceptive BECD score of the AoS patients ranges from 5 to 15, with a mean score of 9.1. The perceptive BECD score of the dysarthria patients ranges from 1 to 12, with a mean score of 6.8. Since there is a mismatch in the perceptive BECD score between the two groups of patients, a similar regression approach as for the age mismatch is considered to validate the classification results (cf. Section 3.4).

The database contains recordings of 8 different sentences at a sampling frequency of 44.1 kHz. After downsampling to 16 kHz and manually removing non-speech segments at the beginning and end of each sentence, all sentences are concatenated and used to extract features for each patient. The average duration of the considered speech material for the AOS and dysarthria patients is 146 s and 75 s respectively.

3.2. Feature extraction

Proposed features. The features proposed in Section 2 consist of the formant frequencies, duration of continuous voiced segments, number of loudness peaks per second, LTAS, and temporal sparsity. The formant frequencies, duration of continuous voiced segments, and number of loudness peaks per second are extracted using [18]. To compute the LTAS and temporal sparsity, signals are first transformed to the frequency domain using a weighted overlap-add short time Fourier transform framework.

Baseline features. To the best of our knowledge, the automatic discrimination between AoS and dysarthria has never been considered in the state-of-the-art literature. Consequently, comparing the proposed approach to state-of-the-art approaches is not possible. For completeness however, the proposed approach is compared to using an SVM with the openSMILE feature set [18]. The openSMILE feature set is an 6473-dimensional vector consisting of functionals of low-level descriptors such as loudness, MFCCs, spectral frequencies, fundamental frequency, or formant frequencies. In an effort to minimize the effects of the curse of dimensionality, we use principal component analyses (PCA) to reduce the dimension of the openSMILE feature set by retaining the PCA features that explain 95% of data variance. Although not presented here due to space constraints, using openSMILE features without PCA yields a significantly worse performance.

3.3. Classification analyses

The used classifiers are SVMs with a radial basis kernel function. Given the small number of patients currently available in the corpus, validation is done following a leave-one-speaker-out validation strategy. For each SVM, features are normalized using the mean and standard deviation of the training data in each fold. The performance is evaluated in terms of the accuracy Acc, i.e., percentage of correctly classified patients. In addition, the percentage of correctly classified AoS patients Acc_{A} and the percentage of correctly classified dysarthria patients Acc_{D} is presented, with $Acc = \frac{Acc_{\text{A}} + Acc_{\text{D}}}{2}$. To select the soft margin constant C and the kernel width γ for the SVMs, nested cross-validation is performed on the training data in each fold with $C \in \{10^{-2}, 10^{4}\}$ and $\gamma \in \{10^{-4}, 10^{2}\}$. The final hyper-parameters used in each fold are selected as the ones resulting in the highest mean accuracy on the training data. When using the openSMILE feature set, the PCA components are learned based on the training data in each fold.

3.4. Regression analyses

As described in Section 3.1, there exists a mismatch in age and perceptive BECD score between the AoS and dysarthria patients. As in [29], to determine whether the presented classification results of the proposed approach are biased by the age or perceptive BECD score of the patients (i.e., to determine whether the proposed feature sets characterize the age or perceptive BECD score of the patients instead of the MSD), we train Support Vector Regressors (SVR) with a Gaussian kernel on each of the three proposed feature sets. The regressors are trained to predict the age or the perceptive BECD score of the patients within a leave-one-speaker-out validation framework. Features are normalized using the mean and standard deviation of the training data in each fold. The performance is evaluated in terms of the coefficient of determination R^2, which assesses how well a model explains and predicts the target variable (i.e., age or perceptive BECD score). A small R^2 (i.e., negative or close to 0) indicates failure to accurately model the data whereas a value of R^2 close to 1 indicates accurate modeling of the data. For completeness, the regression performance is additionally evaluated using the mean absolute error (MAE) between the predicted values and the target values (i.e., age or perceptive BECD score). To select the soft margin constant C and the kernel width γ for the SVRs, nested cross-validation is performed on the training data in each fold with $C \in \{10^{-2}, 10^{4}\}$ and $\gamma \in \{10^{-4}, 10^{2}\}$. The final hyper-parameters are selected as the ones resulting in the highest coefficient of determination R^2 on the training data.

4. Results

Table 1 presents the performance of an SVM using the baseline openSMILE feature set, segmental distortion features (SVM_1), loudness and hyper-nasality features (SVM_2), and syllabification features (SVM_3). In addition, the performance of the pro-

[2] BECD is the French acronym for "Batterie d'Évaluation Clinique de la Dysarthrie" which stands for "Clinical Assessment Test for Dysarthria".

Table 1: *Classification accuracy of an SVM using the baseline openSMILE feature set, segmental distortion features (SVM$_1$), loudness and hyper-nasality features (SVM$_2$), and syllabification features (SVM$_3$), and the classification accuracy of the proposed final MV combination scheme on SVM$_1$, SVM$_2$, and SVM$_3$.*

	openSMILE	SVM$_1$	SVM$_2$	SVM$_3$	MV
Acc [%]	65	85	75	80	**90**
Acc_{A} [%]	80	80	90	70	**100**
Acc_{D} [%]	50	90	60	90	**80**

Table 2: *Regression performance in terms of coefficient of determination R^2 and MAE of an SVR using segmental distortion features (SVR$_1$), loudness and hyper-nasality features (SVR$_2$), and syllabification features (SVR$_3$) trained to predict the age and perceptive BECD score of patients.*

	Age			**Perceptive BECD**		
	SVR$_1$	SVR$_2$	SVR$_3$	SVR$_1$	SVR$_2$	SVR$_3$
R^2	-1.1	-0.3	-0.2	-2.1	-0.2	-6.0
MAE	16.6	13.3	11.3	4.5	3.0	6.9

posed final MV combination scheme on SVM$_1$, SVM$_2$, and SVM$_3$ is presented. Several observations can be made based on the presented results.

First, it can be observed that the accuracy of any of the proposed feature sets (i.e., SVM$_1$, SVM$_2$, or SVM$_3$) and of the final combination scheme MV is higher than the accuracy of the baseline feature set openSMILE. While the discrimination accuracy for AoS patients using openSMILE is 80%, the discrimination accuracy for dysarthria patients is only at a chance level of 50%. These results confirm the advantages of using carefully handcrafted features motivated by clinical-perceptual signs for AoS and dysarthria discrimination. Second, it can be observed that the proposed final combination scheme MV is advantageous and yields a high accuracy of 90%, outperforming the accuracy of the individual classifiers SVM$_1$, SVM$_2$, and SVM$_3$. While the discrimination accuracy for AoS patients using the proposed MV scheme is 100%, the discrimination accuracy for dysarthria patients is 80%. This lower discrimination accuracy of the MV scheme for dysarthria patients can be partly attributed to the lower Acc_{D} achieved by SVM$_2$. Third, comparing the performance of SVM$_1$, SVM$_2$, and SVM$_3$, it can be observed that using the proposed segmental distortion features (i.e., SVM$_1$) yields the highest accuracy of 85%, while the lowest accuracy of 75% is achieved using the proposed loudness and hyper-nasality features (i.e., SVM$_2$). Determining whether the reason behind the lower accuracy of SVM$_2$ lies in the characterization power of loudness variation and hyper-nasality of the used features or in characteristics of the patients in the considered database remains to be investigated in the future.

In summary, the presented results show that the proposed automatic tool for the discrimination of AoS and dysarthria patients is very advantageous. To verify that the presented advantageous results are not biased by the mismatch in age or perceptive BECD score of the patients in the considered database, we train SVRs aiming to predict the age and perceptive BECD score of patients using each of the three proposed feature sets (cf Section 3.4). The first SVR (SVR$_1$) operates on the 10-dimensional feature vector characterizing segmental distortions, the second SVR (SVR$_2$) operates on the 10-dimensional feature vector characterizing loudness and hyper-nasality, and the third SVR (SVR$_3$) operates on the 4-dimensional feature vector characterizing syllabification. Table 2 shows the regression performance in terms of R^2 and MAE for age and perceptive BECD score prediction. It can be observed that for both age and perceptive BECD score prediction and for all feature sets, the coefficient of determination R^2 is negative. These results indicate that the proposed feature sets fail to accurately model the age and perceptive BECD score of the patients. This conclusion is further supported by the high MAE values for both age and perceptive BECD score prediction and for all feature sets. Hence, it can be said that the advantageous classification results of the proposed feature sets (and consequently, of the final MV combination scheme) presented in Table 1 are not biased by the mismatch in age or perceptive BECD score of the patients. Instead, the proposed feature sets characterize signs of the respective MSDs.

In summary, the presented results demonstrate that the proposed approach can be a viable tool to assist clinicians in discriminating between AoS and dysarthria. Its robustness and applicability in more extensive databases needs to be further investigated. In addition, handcrafting more powerful features to characterize loudness variation and hyper-nasality and investigating more powerful combination schemes of the different classifiers remain topics for future research.

5. Conclusion

In this paper, an automatic method to discriminate between AoS and dysarthria using a small number of hand-crafted features has been proposed. To characterize differences in segmental distortions, it has been proposed to use the first formant frequency, the second formant frequency, and the duration of continuous voiced regions. To characterize differences in loudness and hyper-nasality, it has been proposed to use the number of loudness peaks per second and the LTAS. To characterize syllabification differences, it has been proposed to use the temporal sparsity of spectral coefficients. These different feature sets are used to separately train SVMs that capture different clinical-perceptual signs. At test time, the decisions of the different SVMs are combined through a majority voting scheme. Experimental results on a French database of patients suffering from AoS and dysarthria show the applicability and complementary nature of the proposed acoustic features, with the proposed classification approach yielding a high accuracy of 90%.

6. Acknowledgments

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

7. References

[1] J. R. Duffy, *Motor speech disorders: substrates, differential diagnosis, and management*. Missouri, USA: Elsevier Mosby, 2003.

[2] K. A. Josephs, J. R. Duffy, E. A. Strand, M. M. Machulda, M. L. Senjem, A. V. Master, V. J. Lowe, C. R. Jack, and J. L. Whitwell, "Characterizing a neurodegenerative syndrome: primary progressive apraxia of speech," *Brain*, vol. 135, no. 5, pp. 1522–1536, Mar. 2012.

[3] J. R. Duffy, *Parkinson's disease and movement disorders: diagnosis and treatment guidelines for the practicing physician*. New Jersey, USA: Humana Press, 2000, ch. Motor speech disorders: Clues to neurologic diagnosis, pp. 35–53.

[4] W. Ziegler, I. Aichert, and A. Staiger, "Apraxia of speech: concepts and controversies," *Journal of Speech, Language, and Hearing Research*, vol. 55, no. 5, pp. 1485–1501, Oct. 2012.

[5] S. Fonville, H. B. van der Worp, P. Maat, M. Aldenhoven, A. Algra, and J. van Gijn, "Accuracy and inter-observer variation in the classification of dysarthria from speech recordings," *IEEE Transactions on Speech and Audio Processing*, vol. 255, no. 10, pp. 1545–1548, Oct. 2008.

[6] K. Bunton, R. Kent, J. R. Duffy, J. Rosenbek, and J. Kent, "Listener agreement for auditory-perceptual ratings of dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 6, pp. 1481–1495, Jan. 2008.

[7] K. Haley, A. Jacks, J. Richardson, and J. Wambaugh, "Perceptually salient sound distortions and apraxia of speech: A performance continuum," *American Journal of Speech-Language Pathology*, vol. 26, no. 2S, pp. 631–640, Jun. 2017.

[8] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, May 2012.

[9] J. R. Orozco-Arroyave, F. Hönig, J. Arias-Londoño, J. Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinson's disease," in *Proc. Annual Conference of the International Speech Communication Association*, Dresden, Germany, Sep. 2015, pp. 95–99.

[10] D. Hemmerling, J. R. Orozco-Arroyave, A. Skalski, J. Gajda, and E. Nöth, "Automatic detection of Parkinson's disease based on modulated vowels," in *Proc. Annual Conference of the International Speech Communication Association*, San Francisco, USA, Sep. 2016, pp. 1190–1194.

[11] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech," *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 1, pp. 114–125, Feb. 2010.

[12] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, and V. Uloza, "Exploring similarity-based classification of larynx disorders from human voice," *Speech Communication*, vol. 54, no. 5, pp. 601–610, Jun. 2012.

[13] R. Behroozmand and F. Almasganj, "Comparison of Neural Networks and Support Vector Machines applied to optimized features extracted from patients' speech signal for classification of vocal fold inflammation," in *Proc. IEEE International Symposium on Signal Processing and Information Technology*, Athens, Greece, Jan. 2006, pp. 844–849.

[14] A. A. Dibazar, T. W. Berger, and S. S. Narayanan, "Pathological voice assessment," in *Proc. International Conference of the IEEE Engineering in Medicine and Biology Society*, New York, USA, Sep. 2006, pp. 1669–1673.

[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] T. Bocklet, S. Steidl, E. Nöth, and S. Skodda, "Automatic evaluation of Parkinson's speech - acoustic, prosodic and voice related cues," in *Proc. 14th Annual Conference of the International Speech Communication Association*, Lyon, France, Sep. 2013, pp. 1149–1153.

[17] R. Norel, M. Pietrowicz, C. Agurto, S. Rishoni, and G. Cecchi, "Detection of Amyotrophic Lateral Sclerosis (ALS) via acoustic analysis," in *Proc. 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, Sep. 2018, pp. 377–381.

[18] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. ACM Multimedia*, Barcelona, Spain, Oct. 2018, pp. 835–838.

[19] W. Ziegler and D. von Cramon, "Disturbed coarticulation in apraxia of speech: acoustic evidence," *Brain and Language*, vol. 29, no. 1, pp. 34–47, Sep. 1986.

[20] S. P. Whiteside and R. A. Varley, "A reconceptualisation of apraxia of speech: a synthesis of evidence," *Cortex*, vol. 34, no. 2, pp. 221–231, Apr. 1998.

[21] D. B. den Ouden, E. Galkina, A. Basilakos, and J. Fridriksson, "Vowel formant dispersion reflects severity of apraxia of speech," *Aphasiology*, vol. 32, no. 8, pp. 902–921, Oct. 2018.

[22] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, Apr. 2016.

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

[24] V. Berisha, R. Utianski, and J. Liss, "Towards a clinical tool for automatic intelligibility assessment," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013, pp. 2825–2828.

[25] I. Kodrasi and H. Bourlard, "Statistical modeling of speech spectral coefficients in patients with Parkinson's disease," in *Proc. ITG conference on Speech Communication*, Oldenburg, Germany, Oct. 2018, pp. 271–275.

[26] ——, "Super-Gaussianity of speech spectral coefficients as a potential biomarker for dysarthric speech detection," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brighton, UK, May 2019, pp. 6400–6404.

[27] ——, "Spectro-temporal sparsity characterization for dysarthric speech detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 1, pp. 1210–1222, Dec. 2020.

[28] P. Auzou and V. Rolland-Monnoury, *Batterie d'évaluation clinique de la dysarthrie*. Isbergues, France: Ortho Édition, 2006.

[29] T. Arias-Vergara, J. R. Orozco-Arroyave, M. Cernak, S. Gollwitzer, M. Schuster, and E. Nöth, "Phone-attribute posteriors to evaluate the speech of cochlear implant users," in *Proc. Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep. 2019, pp. 3108–3112.