Spectro-Temporal Sparsity Characterization for Dysarthric Speech Detection

Ina Kodrasi, *Member, IEEE*, and Hervé Bourlard, *Fellow, IEEE*

Abstract*—To assist the clinical diagnosis and treatment of neurological diseases that cause speech dysarthria such as Parkinson's disease (PD), it is of paramount importance to craft robust features which can be used to automatically discriminate between healthy and dysarthric speech. Since dysarthric speech of patients suffering from PD is breathy, semi-whispery, and is characterized by abnormal pauses and imprecise articulation, it can be expected that its spectro-temporal sparsity differs from the spectro-temporal sparsity of healthy speech. While we have recently successfully used temporal sparsity characterization for dysarthric speech detection, characterizing spectral sparsity poses the challenge of constructing a valid feature vector from signals with a different number of unaligned time frames. Further, although several non-parametric and parametric measures of sparsity exist, it is unknown which sparsity measure yields the best performance in the context of dysarthric speech detection. The objective of this paper is to demonstrate the advantages of spectro-temporal sparsity characterization for automatic dysarthric speech detection. To this end, we first provide a numerical analysis of the suitability of different non-parametric and parametric measures (i.e., l_1-norm, kurtosis, Shannon entropy, Gini index, shape parameter of a Chi distribution, and shape parameter of a Weibull distribution) for sparsity characterization. It is shown that kurtosis, the Gini index, and the parametric sparsity measures are advantageous sparsity measures, whereas the l_1-norm and entropy measures fail to robustly characterize the temporal sparsity of signals with a different number of time frames. Second, we propose to characterize the spectral sparsity of an utterance by initially time-aligning it to the same utterance uttered by a (arbitrarily selected) reference speaker using dynamic time warping. Experimental results on a Spanish database of healthy and dysarthric speech show that estimating the spectro-temporal sparsity using the Gini index or the parametric sparsity measures and using it as a feature in a support vector machine results in a high classification accuracy of 83.3%.

***Index Terms*—Non-parametric sparsity, parametric sparsity, SVM, DTW, Parkinson's disease.**

I. Introduction

BECAUSE of increasing population numbers and aging, the prevalence of neurological disorders such as Parkinson's disease (PD) is also increasing [1]. The number of people requiring screening and treatment will continue to grow in the coming decades, likely putting a strain on the health care system. Besides other motor and non-motor symptoms, many PD patients develop speech dysarthria, which is an impairment that affects several components of the speech production mechanism such as phonation, articulation, and prosody [2].

To diagnose neurological disorders such as PD and evaluate their progression, clinicians exploit several examinations which assess different motor and sensory skills. However, such examinations can be subject to the expertise of the clinician and might be affected by their familiarity with the patient. Aiming to assist the clinical diagnosis and treatment of patients suffering from neurological disorders, there has been a growing interest in the research community to develop discriminatory features which can be used for the automatic detection and monitoring of dysarthric speech.

To quantify impacted phonation and characterize disturbances in vocal fold vibration as well as excessive turbulence due to incomplete closure of the vocal folds, features such as fundamental frequency, jitter, shimmer, or harmonics-to-noise ratio (HNR) have been used [3]–[6]. To quantify impacted articulation and characterize vocal tract atypicalities, features such as Mel frequency cepstral coefficients (MFCCs), linear prediction coefficients (LPCs), and perceptual LPCs have been used [6]–[9]. Impacted articulation has also been characterized using features such as vowel space area, vowel articulation index, consonant spectral trend, or formant centralization ratio [6], [10]–[13]. Segment-dependent changes in different speech production components have also been characterized through capturing changes in phoneme duration, frequencies, pitch, and formant slopes [14]. Recently, we have proposed to jointly quantify impacted phonation and articulation by characterizing the temporal sparsity of the speech spectral coefficients [15], [16]. Temporal sparsity refers to the sparsity (i.e., lack of energy) of coefficients in a single subband across different time frames and arises due to, e.g., pauses between phonemes. In [15], [16] we have shown that because of temporal smearing from imprecise articulation, breathiness, and abnormal pauses, dysarthric speech spectral coefficients are less temporally sparse than healthy speech spectral coefficients. Further, it is shown that a support vector machine (SVM) achieves a higher accuracy for healthy and dysarthric speech detection when the feature vector is constructed from temporal sparsity estimates rather than from commonly used features such as fundamental frequency, jitter, shimmer, or HNR [16].

While temporal sparsity can indeed be a powerful discriminator between healthy and dysarthric speech, spectral sparsity

Manuscript received August 14, 2019; revised January 30, 2020 and March 30, 2020; accepted March 30, 2020. Date of publication April 6, 2020; date of current version April 23, 2020. This work was supported by 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". The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Carlos Busso. *(Corresponding author: Ina Kodrasi.)*

The authors are with the Idiap Research Institute, 1920 Martigny, Switzerland (e-mail: ina.kodrasi@idiap.ch; herve.bourlard@idiap.ch).

Digital Object Identifier 10.1109/TASLP.2020.2985066

has never been exploited. It can be expected that impacted phonation, articulation, and respiration do not only affect the temporal sparsity of dysarthric speech signals but also their spectral sparsity. However, spectral sparsity is a time-dependent feature, additionally reflecting what is being uttered in each time frame. Due to different speakers and speaking rates, utterances uttered by healthy speakers and speakers with dysarthria are unaligned and of different length. Hence, assessing and using spectral sparsity as a discriminative feature between healthy and dysarthric speech poses the challenge of constructing a valid feature vector from speech signals with a different number of unaligned time frames.

Further, both for temporal and spectral sparsity characterization, a natural question that arises is how to assess sparsity. Although sparse representations arise in numerous areas such as image processing [17], speech signal processing [18], [19], computer vision [20], and pattern recognition [21], sparsity is not uniquely defined and several non-parametric and parametric measures of sparsity have been used in the literature. Commonly used non-parametric sparsity measures are the l_1-norm, the kurtosis, the Shannon entropy, and the Gini index [22]. Commonly used parametric sparsity measures are the shape parameters of a Chi or Weibull distribution [16], [23], [24]. To the best of our knowledge, the suitability of these different sparsity measures for dysarthric speech detection has never been investigated.

The objective of this paper is to demonstrate the advantages of spectro-temporal sparsity characterization for automatic dysarthric speech detection. To this end, we first provide a numerical analysis of the applicability of different non-parametric and parametric measures, i.e., l_1-norm, kurtosis, Shannon entropy, Gini index, shape parameter of a Chi distribution, and shape parameter of a Weibull distribution, for spectro-temporal sparsity characterization. We show that kurtosis, the Gini index, and the parametric sparsity measures can accurately characterize sparsity, whereas the l_1-norm and entropy measures fail to robustly characterize the sparsity of signals with a different number of time frames. Second, we propose to characterize the spectral sparsity of an utterance by initially time-aligning it to the same utterance uttered by a (arbitrary selected) reference speaker by means of dynamic time warping (DTW) [25]. Spectral sparsity can then be estimated for each time frame and the spectral sparsity feature vector can be constructed by concatenating the sparsity estimates for all time frames. Using a Spanish database of healthy and dysarthric speech, we show that compared to temporal sparsity, spectral sparsity is a more powerful discriminator for dysarthric speech detection. Additionally, it is shown that exploiting both the temporal and spectral sparsity characterization, an even higher classification accuracy can be achieved. Out of the considered sparsity measures, the Gini index and the parametric sparsity measures yield the highest performance when using spectro-temporal sparsity as a feature vector in an SVM for healthy and dysarthric speech detection.

The paper is organized as follows. In Section II we provide insights on the spectro-temporal sparsity of healthy and dysarthric speech signals. In Section III we define the considered non-parametric and parametric measures of sparsity. Section IV presents a numerical analysis of the suitability of the considered sparsity measures for comparing the spectro-temporal sparsity of speech signals. Section V presents the method proposed to align signals and construct spectral sparsity characterization features. In Section VI we present experimental results evaluating the applicability of the considered sparsity measures for dysarthric speech detection.

II. Spectro-Temporal Sparsity of Healthy and Dysarthric Speech

Supported by empirical observations, e.g., in [24], [26], [27], it is widely accepted that speech spectral coefficients are spectro-temporally sparse.[1] On the one hand, spectral sparsity refers to the sparsity of coefficients in a single time frame across different subbands and arises due to, e.g., formant transitions in voiced sounds. On the other hand, temporal sparsity refers to the sparsity of coefficients in a single subband across different time frames and arises due to, e.g., pauses between phonemes. In [15], [16] we have shown that because of temporal smearing from imprecise articulation, breathiness, and abnormal pauses, dysarthric speech spectral coefficients are less temporally sparse than healthy speech spectral coefficients. However, it can be expected that imprecise articulation, breathiness, abnormal pauses, and vocal tremor also affect the spectral sparsity of dysarthric speech spectral coefficients.

Fig. 1 depicts the spectrograms of the same utterance uttered by a healthy speaker and a speaker suffering from PD. In addition, bounding boxes highlighting the spectral coefficients that are used to compute the temporal sparsity at the exemplary subband index $k = 50$ and the spectral sparsity at the exemplary time frame index $l = 20$ are also depicted. Several observations can be made for the depicted exemplary spectrograms. First, it can be observed that speech spectral coefficients are indeed spectro-temporally sparse, independently of whether healthy or dysarthric speech is considered. In both spectrograms, speech is present only in some time frames, and in these time frames, only some subbands have significant energy while several subbands have (nearly) no energy. Second, the depicted spectrograms show that the dysarthric speech signal is smeared when compared to the healthy speech signal, with energy in time-frequency bins where the healthy speech signal has no energy (cf. e.g., the spectral coefficients at subband index $k = 50$). Such smearing should yield a difference in the spectro-temporal sparsity of the healthy and dysarthric speech spectral coefficients. Third, the depicted spectrograms show that to compare the temporal sparsity of healthy and dysarthric speech in each subband, vectors of spectral coefficients of different lengths need to be taken into account. In these exemplary spectrograms, the healthy utterance has a length of $L = 40$ time frames, whereas the

[1] In this paper, spectro-temporal sparsity refers to the fact that speech is present only in some time frames, and in these time frames, only some subbands have significant energy while several subbands have (nearly) no energy. It should be noted that establishing a theoretically solid definition of how many time frames and subbands should lack energy for the speech spectral coefficients to be considered sparse is beyond the scope of this paper. Instead, our objective is to establish an accurate characterization of the difference in the number of time frames and subbands lacking energy for healthy and dysarthric speech spectral coefficients.

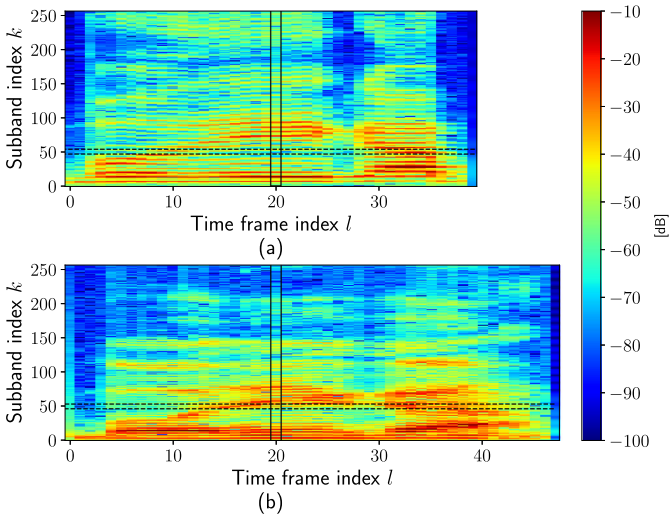

Fig. 1. Spectrogram of the same utterance uttered by a (a) healthy speaker and (b) speaker suffering from PD. The dashed bounding box highlights the coefficients to consider when computing the temporal sparsity at subband index $k = 50$. The solid bounding box highlights the coefficients to consider when computing the spectral sparsity at time frame index $l = 20$. It can be observed that speech spectral coefficients are spectro-temporally sparse, with many time frames and subbands having (nearly) no energy. Further, it can be observed that the spectro-temporal sparsity of healthy and dysarthric speech spectral coefficients differs. To compare the temporal sparsity of healthy and dysarthric speech in each subband, vectors of different lengths need to be considered, whereas to compare the spectral sparsity, time frames corresponding to the same phonetic content need to be considered.

dysarthric utterance has a length of $L = 48$ time frames. Finally, the depicted spectrograms show that spectral sparsity is a time-dependent feature. In these exemplary spectrograms, comparing the spectral sparsity at time frame index $l = 20$ is meaningless, since signals are not aligned and the spectral coefficients at this time frame might represent different phonetic content for the healthy and dysarthric speech. To compare the spectral sparsity of healthy and dysarthric speech, time frames corresponding to the same phonetic content need to be considered.

III. Non-Parametric and Parametric Sparsity Measures

In this section, the considered non-parametric and parametric measures of sparsity are defined. Speech spectral coefficients are denoted by $S_{k,l}$, with k the subband index and l the time frame index. In addition, the vector of spectral magnitudes in subband k is denoted by

$$\mathbf{a}_k = [|S_{k,1}|\ |S_{k,2}|\ \cdots\ |S_{k,L}|]^T, \tag{1}$$

with L being the total number of time frames. Similarly, the vector of spectral magnitudes in time frame l is denoted by

$$\mathbf{a}_l = [|S_{1,l}|\ |S_{2,l}|\ \cdots\ |S_{K,l}|]^T, \tag{2}$$

with K being the total number of subbands. Furthermore, the I-dimensional vector $\mathbf{a} = [a_1\ a_2\ \cdots\ a_I]^T$ is used to refer to any vector $\mathbf{a}_k$, $k = 1, \ldots, K$ and $\mathbf{a}_l$, $l = 1, \ldots, L$.

A. Non-Parametric Sparsity Measures

The most commonly investigated non-parametric sparsity measures are the l_p-norm measures, with $0 \leq p \leq 1$ [22]. While the l_0-norm is the traditional sparsity measure, it is sensitive to noise and unsuited in practice to be used in applications where signal sparsity is the desired outcome. The l_0-norm is non-convex and optimization problems with non-convex penalty functions are typically hard (if not impossible) to solve, particularly for large scale problems [28]. A commonly used convex relaxation of the l_0-norm is the l_1-norm, defined as [29]

$$l_1(\mathbf{a}) = \sum_{i=1}^{I} a_i. \tag{3}$$

Another popular non-parametric sparsity measure is the kurtosis κ [30], [31], which characterizes whether the data is light- or heavy-tailed relative to the normal distribution. The kurtosis is defined as

$$\kappa(\mathbf{a}) = \frac{\sum_{i=1}^{I} a_i^4}{\left(\sum_{i=1}^{I} a_i^2\right)^2}. \tag{4}$$

Rao and Kreutz-Delgado use entropy measures such as the Shannon entropy ϵ as non-parametric measures of sparsity [32]. The Shannon entropy is defined as

$$\epsilon(\mathbf{a}) = -\sum_{i=1}^{I} \tilde{a}_i \ln \tilde{a}_i, \tag{5}$$

with $\tilde{a}_i = \frac{a_i^2}{\|\mathbf{a}\|_2^2}$. Finally, Hurley and Rickard have shown that an advantageous non-parametric sparsity measure is the Gini index G, originally proposed in economics as a measure of wealth inequality [22], [33]. The Gini index is defined as

$$G(\mathbf{a}) = 1 - 2\sum_{i=1}^{I} \frac{a_{(i)}}{l_1(\mathbf{a})}\left(\frac{I - i + \frac{1}{2}}{I}\right), \tag{6}$$

with $a_{(i)}$ denoting spectral magnitudes ordered in ascending order, i.e., $a_{(1)} \leq a_{(2)} \leq \cdots \leq a_{(I)}$. Hurley and Rickard have thoroughly analyzed the above-mentioned non-parametric sparsity measures, showing that only the Gini index satisfies all properties that an advantageous sparsity measure should have (such as (in)sensitivity to multiplicative constants, additive constants, or data length) [22]. Nevertheless, the Gini index is typically not used in the speech and audio research community to compare the sparsity of different signals, with l_p-norm measures and kurtosis being used instead.

B. Parametric Sparsity Measures

In contrast to non-parametric measures, sparsity can also be assessed by parametric measures such as the shape parameter of a Chi distribution or the shape parameter of a Weibull distribution [16], [23], [24], [27]. In [24], [27], empirical analyses show that the Chi and Weibull distributions model histograms of speech spectral amplitudes with a high accuracy. Such models hold both locally, i.e., when observing the distribution of speech spectral amplitudes in a single time-frequency bin, as well

as globally, i.e., when considering the distribution of speech spectral amplitudes in a single subband across time. Assuming that the global distribution of speech spectral amplitudes can be well-modeled with a Chi or Weibull distribution, a maximum likelihood (ML) procedure can be used to estimate the distribution shape parameter that characterizes sparsity, with lower shape parameter values describing more sparse data. Differently from the well-investigated non-parametric sparsity measures, to the best of our knowledge these parametric measures of sparsity have not been thoroughly analyzed with respect to, e.g., their sensitivity to data length or to erroneous assumptions about the data distribution. In the remainder of this section, the ML estimation procedures of the shape parameter of the Chi and Weibull distributions are presented.

The density function of the Chi distribution is given by

$$p_{\mathrm{c}}(a) = \frac{2}{\Gamma(\mu)} \left(\frac{\mu}{\sigma^2}\right)^{\mu} a^{2\mu-1} e^{-\frac{\mu}{\sigma^2}a^2}, \tag{7}$$

with $\sigma^2 = \mathcal{E}\{a^2\}$, μ the shape parameter, and $\Gamma(\cdot)$ the complete Gamma function. For $\mu < 1$, the Chi distribution in (7) models sparse spectral coefficients, with lower values of μ corresponding to more sparse spectral coefficients [23], [24]. Given the spectral magnitudes in $\mathbf{a}$, the shape parameter μ can be estimated using the following ML estimation procedure. Ignoring the terms independent of μ, the log-likelihood function is given by

$$\ln \mathcal{L}_{\mathrm{C}}(\mu) = -I \ln \Gamma(\mu) + I\mu \ln \mu - I\mu \ln \sigma^2 + (2\mu - 1) \sum_{i=1}^{I} \ln a_i - \frac{\mu}{\sigma^2} \sum_{i=1}^{I} a_i^2. \tag{8}$$

To compute an ML estimate of the shape parameter μ, the negative of the log-likelihood function in (8) should be minimized. Since no analytical solution can be found, an iterative optimization technique such as the Newton method can be used [34]. To improve the robustness and convergence speed, the analytical gradient of the negative log-likelihood function can be used in the optimization routine, i.e.,

$$\frac{\partial[-\ln \mathcal{L}_{\mathrm{C}}(\mu)]}{\partial \mu} = I\psi(\mu) - I - I \ln \mu + I \ln \sigma^2 - 2 \sum_{i=1}^{I} \ln a_i + \frac{1}{\sigma^2} \sum_{i=1}^{I} a_i^2, \tag{9}$$

with $\psi(\cdot)$ the digamma function.

The density function of the Weibull distribution is given by

$$p_{\mathrm{w}}(a) = \frac{\theta}{\xi} \left(\frac{a}{\xi}\right)^{\theta-1} e^{-\left(\frac{a}{\xi}\right)^{\theta}}, \tag{10}$$

with θ the shape parameter and ξ the scale parameter which can be expressed as

$$\xi = \frac{\sigma}{\sqrt{\Gamma\left(1 + \frac{2}{\theta}\right)}}. \tag{11}$$

Given the spectral magnitudes in $\mathbf{a}$, the shape parameter θ can be estimated using the following ML estimation procedure. The log-likelihood function of θ is given by

$$\ln \mathcal{L}_{\mathrm{W}}(\theta) = I \ln \theta - I\theta \ln \xi + (\theta - 1) \sum_{i=1}^{I} \ln a_i - \frac{1}{\xi^{\theta}} \sum_{i=1}^{I} a_i^{\theta}. \tag{12}$$

To compute an ML estimate of the shape parameter θ, the negative of the log-likelihood function in (12) should be minimized with an iterative optimization technique. To improve the robustness and convergence speed of the optimization routine, the gradient of the negative log-likelihood function can also be provided, i.e.,

$$\frac{\partial[-\ln \mathcal{L}_{\mathrm{W}}(\theta)]}{\partial \theta} = -\frac{I}{\theta} + I \ln \sigma - \frac{I}{2}\eta + \frac{I}{\theta}\lambda - \sum_{i=1}^{I} \ln a_i + \frac{1}{\xi^{\theta}} \left(\frac{1}{2}\eta - \frac{1}{\theta}\lambda - \ln \sigma\right) \sum_{i=1}^{I} a_i^{\theta} + \frac{1}{\xi^{\theta}} \sum_{i=1}^{I} a_i^{\theta} \ln a_i, \tag{13}$$

where the variables $\eta = \ln \Gamma(1 + \frac{2}{\theta})$ and $\lambda = \psi(1 + \frac{2}{\theta})$ have been defined for ease of notation.

IV. Numerical Analysis of Sparsity Measures

Similarly to the analysis presented in [22], in this section we investigate the suitability of all considered measures to assess the sparsity of data drawn from a set of parameterized distributions. Differently from [22], we do not only consider non-parametric sparsity measures, but also parametric sparsity measures. Furthermore, since our end goal is to compare the sparsity of speech spectral coefficients, the considered parameterized distributions differ from [22].

When comparing the sparsity of vectors of spectral coefficients from different speech signals, we would like the used sparsity measure to have the following properties.

i) Clearly, the used measure should reflect differences in the sparsity between the vectors, i.e., for the vector that is more sparse, the l_1-norm, the Shannon entropy, and the estimated shape parameters of the Chi and Weibull distributions should be smaller, whereas the kurtosis and the Gini index should be larger (cf. the definition of the sparsity measures in Section III).

ii) As described in Section II, when comparing the temporal sparsity of vectors of spectral coefficients from different speech signals, the vectors will typically be of different lengths. Hence, the used sparsity measure should be insensitive to data length, i.e., when a vector is extended with coefficients with the same sparsity, the sparsity measure should not change.

iii) For the parametric sparsity measures, i) and ii) should hold also when the true distribution of the speech spectral magnitudes deviates from the assumed distributions (i.e., the Chi or Weibull distribution).

To illustrate the behavior of the considered sparsity measures with respect to i), we draw $I = 10000$ coefficients from a Chi

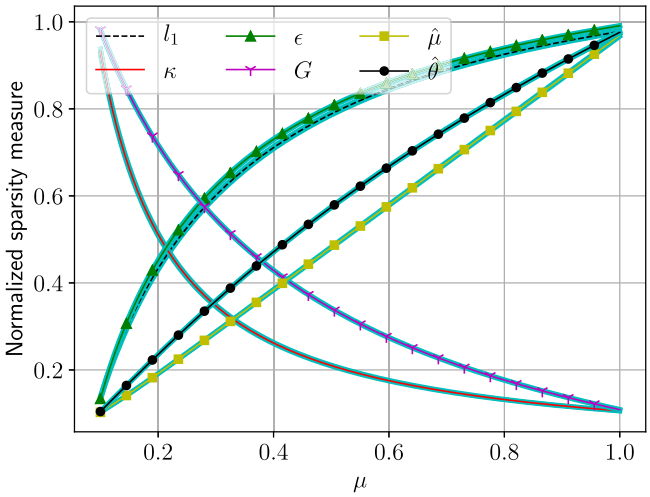

Fig. 2. Sparsity of $I = 10000$ coefficients drawn from a Chi distribution with different values of the shape parameter μ. All measures are scaled between 0.1 and 1. Solid cyan lines for each sparsity measure illustrate the 95% confidence interval for 100 repetitions of drawing coefficients.

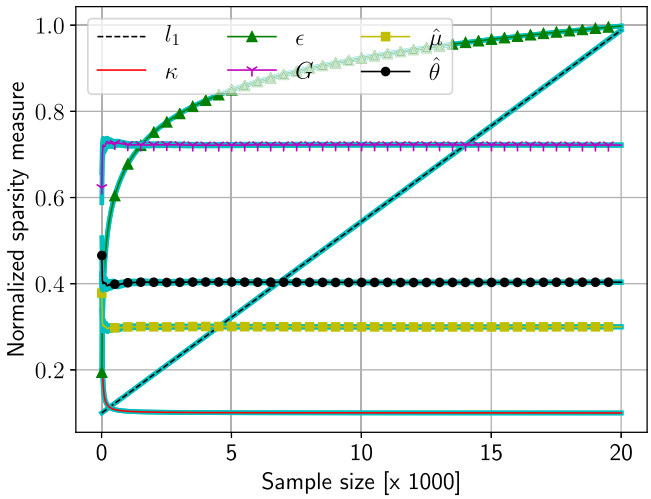

Fig. 3. Sparsity of a variable set of coefficients drawn from a Chi distribution with $\mu = 0.5$. All measures are scaled between 0.1 and 1. Solid cyan lines for each sparsity measure illustrate the 95% confidence interval for 100 repetitions of drawing coefficients.

distribution for different values of μ, i.e., $\mu \in [0.1, 1]$. For each drawn set of coefficients, the different non-parametric and parametric sparsity measures are computed. For each μ, coefficients are drawn 100 times such that the 95% confidence interval for the different sparsity measures can be computed. As μ increases (i.e., as the drawn coefficients become less sparse), the l_1-norm, the Shannon entropy, and the estimated shape parameters for the Chi and Weibull distributions should increase, whereas the kurtosis and the Gini index should decrease. Fig. 2 depicts the computed non-parametric and parametric sparsity measures for the different sets of coefficients. All sparsity measures are scaled between 0.1 and 1 for visual convenience. It can be observed that all considered measures reflect the decrease in sparsity as μ increases. We therefore conclude that all considered sparsity measures satisfy i) if the speech spectral coefficients follow the commonly assumed Chi distribution.

To illustrate the behavior of the considered sparsity measures with respect to ii), we draw a variable set of coefficients $I \in [20, 20000]$ from a Chi distribution with a fixed value of the shape parameter $\mu = 0.5$. For each I, coefficients are drawn 100 times such that the 95% confidence interval for the different sparsity measures can be computed. Since the coefficients are drawn from the same distribution, an advantageous sparsity measure should quickly converge as the number of drawn coefficients increases. Fig. 3 depicts the computed non-parametric and parametric sparsity measures for the different sets of coefficients. All sparsity measures are scaled between 0.1 and 1 for visual convenience. It can be observed that the parametric measures (i.e., $\hat{\mu}$ and $\hat{\theta}$) converge quickly. Further it can be observed that out of the considered non-parametric measures, only the kurtosis κ and the Gini index G converge, with the l_1-norm and the Shannon entropy ϵ yielding a different sparsity assessment as the number of drawn coefficients increases. Since the sparsity measure needs to be insensitive to data length when comparing spectral coefficients from different signals, these results show

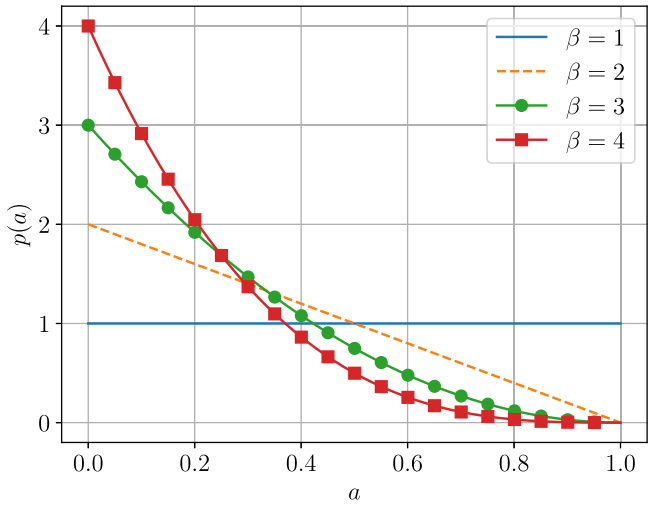

Fig. 4. Density function of the Beta distribution for different values of the shape parameter β.

that the only applicable measures are the kurtosis, the Gini index, and the parametric sparsity measures.

To illustrate the behavior of the considered sparsity measures with respect to iii), we repeat the previously presented numerical analysis using coefficients drawn from a Beta distribution. The density function of the Beta distribution is given by

$$p(a) = \beta(1-a)^{\beta-1}, \quad \text{for} \quad a \in [0, 1], \tag{14}$$

with β denoting the shape parameter. Fig. 4 depicts the density function of the Beta distribution for different values of the shape parameter β. It can be observed that for $\beta = 1$, the uniform distribution is obtained. As β increases, fewer coefficients with a large amplitude are obtained, while the number of coefficients with a small amplitude increases. Hence, when drawing coefficients from a Beta distribution for increasing values of the shape parameter β, the sparsity of the drawn coefficients increases. Being a

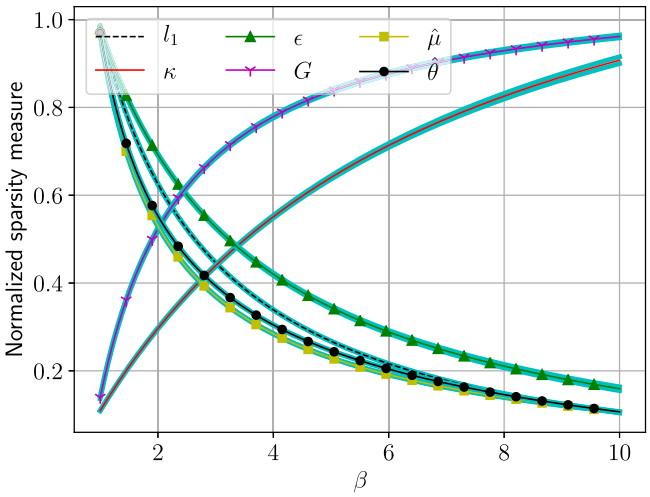

Fig. 5. Sparsity of $I = 10000$ coefficients drawn from a Beta distribution with different values of the shape parameter β. All measures are scaled between 0.1 and 1. Solid cyan lines for each sparsity measure illustrate the 95% confidence interval for 100 repetitions of drawing coefficients.

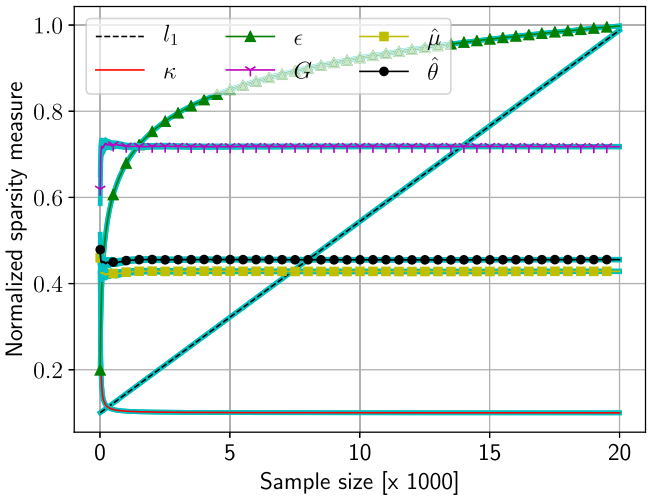

Fig. 6. Sparsity of a variable set of coefficients drawn from a Beta distribution with $\beta = 4$. All measures are scaled between 0.1 and 1. Solid cyan lines for each sparsity measure illustrate the 95% confidence interval for 100 repetitions of drawing coefficients.

bounded distribution, the Beta distribution cannot model speech spectral magnitudes. The objective of the following numerical analysis is to illustrate the behavior of the considered sparsity measures if the distribution of the speech spectral magnitudes differs from the commonly assumed distributions.

The behavior of the considered sparsity measures when assessing the sparsity of coefficients drawn from a Beta distribution is depicted in Figs. 5 and 6. Fig. 5 depicts the normalized sparsity measures for a fixed sample size I and increasing values of β, i.e., $I = 10000$, $\beta \in [1, 10]$, whereas Fig. 6 depicts the normalized sparsity measures for increasing sample size and a fixed value of β, i.e., $I \in [20, 20000]$, $\beta = 4$. For each considered β and I, coefficients are drawn 100 times such that the 95% confidence interval for the different measures of sparsity can be computed. Fig. 5 shows that also when the coefficients are drawn from a Beta distribution, all considered sparsity measures reflect the change in sparsity, i.e., as the sparsity of the coefficients increases, the l_1-norm, the Shannon entropy, and the estimated shape parameters of the Chi and Weibull distributions decrease, whereas the kurtosis and the Gini index increase. Furthermore, Fig. 6 shows that similarly to Fig. 3, only the kurtosis, the Gini index, and the parametric measures of sparsity converge as the sample size increases, whereas the l_1-norm and the Shannon entropy do not converge.

In summary, the presented numerical analyses show that for data generated according to a Chi distribution (which has been empirically shown to model speech spectral amplitudes with a high accuracy) or a Beta distribution (which cannot model speech spectral amplitudes), the kurtosis, the Gini index, and the parametric measures of sparsity are the only measures which can be used to robustly compare sparsity. In Section VI, these measures are used to create the feature vector for an SVM classifying healthy and dysarthric speech. As will be shown in Section VI, characterizing the spectro-temporal sparsity using kurtosis, which has been often investigated as a measure for characterizing pathological speech [31], yields the worst classification accuracy, whereas using the Gini index and the parametric sparsity measures results in a similarly high classification accuracy.

V. Time Alignment for Spectral Sparsity Characterization

As described in Section II, the spectro-temporal sparsity of dysarthric speech spectral coefficients differs from the spectro-temporal sparsity of healthy speech spectral coefficients. In [16] we have shown that creating the temporal sparsity feature vector can be straightforwardly done by computing the spectral magnitudes in each subband, computing the sparsity measure, and concatenating the sparsity estimates for each subband. Characterizing spectral sparsity on the other hand is not straightforward since signals are unaligned and of different length (cf. Fig. 1). In this section we propose to use DTW [25] to align all available signals to the corresponding signals from a (arbitrary selected) reference speaker before estimating the spectral sparsity. It should be noted that since the phonetic content of each utterance is known in advance, the desired alignments can also be obtained using a Hidden Markov Model-based system in forced alignment mode. Since experimental results suggest that DTW already yields a very good alignment performance for our application, DTW is used in this manuscript.

Let $\mathbf{S}_r$ denote the $(L_r \times K)$-dimensional subband representation of an utterance from the reference speaker r, with L_r denoting the total number of time frames. Similarly, let $\mathbf{S}_t$ denote the $(L_t \times K)$-dimensional subband representation of the same utterance from another speaker t, with L_t denoting the total number of time frames. The representations $\mathbf{S}_r$ and $\mathbf{S}_t$ are typically not aligned (due to different speakers and speaking rates) and are generally of different lengths, i.e., $L_r \neq L_t$. We propose to align these two representations through DTW using the logarithmic magnitude and a simple Euclidean distance as the cost function [25]. For each time frame l in $\mathbf{S}_r$, $l = 1, \ldots, L_r$, we extract all time frames in $\mathbf{S}_t$ that are mapped to it by DTW. Each

frame l in the reference utterance is mapped to one or multiple frames in $\mathbf{S}_t$. If the number of time frames in $\mathbf{S}_r$ is larger than the number of time frames in $\mathbf{S}_t$, the same frame in $\mathbf{S}_t$ is mapped to multiple frames in $\mathbf{S}_r$. If the number of time frames in $\mathbf{S}_r$ is smaller than the number of time frames in $\mathbf{S}_t$, multiple frames in $\mathbf{S}_t$ are mapped to the same frame in $\mathbf{S}_r$. The spectral sparsity is individually estimated for each of these extracted time frames and the spectral sparsity of $\mathbf{S}_t$ for time frame l is computed as the average of the spectral sparsity estimates across all extracted time frames. This process is repeated for all available utterances and speakers. Following such a procedure, the dimension of the spectral sparsity feature vector is L_r, i.e., it is dictated by the number of time frames in the subband representation of the utterance from the reference speaker r. In Section VI-D it is shown that the classification performance of a classifier exploiting the so-computed spectral feature vector is insensitive to the selected reference speaker r.

It should be noted that since the characterization of spectral sparsity requires time frames corresponding to the same phonetic content to be considered, using spectral sparsity for healthy and dysarthric speech classification is only applicable when recordings of the same speech material across all speakers are available. Hence, such a feature vector cannot be used in e.g., spontaneous speech tasks.

VI. Results and Discussion

In this section, empirical analyses of the spectro-temporal sparsity of healthy and dysarthric speech caused by PD are presented. The spectro-temporal sparsity is assessed using kurtosis, the Gini index, and the parametric sparsity measures. For each sparsity measure, statistical significance analyses are conducted to compare the mean spectro-temporal sparsity estimates across healthy speakers and PD patients. In addition, the classification accuracy of an SVM aiming at healthy and dysarthric speech classification using the proposed sparsity estimates is investigated and compared to using state-of-the-art features such as fundamental frequency, jitter, shimmer, HNR, and MFCCs.

A. Databases and Preprocessing

We consider Spanish recordings of 45 healthy speakers and 45 PD patients from [35], with all speakers being Colombian Spanish native speakers. Both groups of speakers contain 22 male and 23 female speakers. The age of the healthy speakers ranges from 31 to 86 years old, with a mean age of 61.0 and a standard deviation of 9.7. The age of the PD patients ranges from 33 to 81 years old, with a mean age of 61.0 and a standard deviation of 9.5. All recordings are done in a sound-proof booth using an omnidirectional Shure SM63 L microphone. Hence, the considered database is well-balanced in terms of age and gender and the recording conditions between the two groups of speakers are the same.

All PD patients were diagnosed by neurologist experts, were labeled according to the Unified Parkinson's Disease Rating Scale (UPDRS) scale [36], and were recorded no more than 3 hours after the morning medication. The number of years after diagnosis ranges from 0.4 to 43 years, with a mean of 11.0 and a standard deviation of 9.5. The UPDRS scores range from 13 to 75, with a mean score of 39.1 and a standard deviation of 15.1.

The sampling frequency of all recordings is 44.1 kHz and we consider recordings of 10 words uttered by all speakers. All recordings are downsampled to 16 kHz and manual voice activity detection is performed to extract the speech-only segments. Concatenating the extracted speech-only segments for each speaker yields a signal with a mean length of 6.1 s across the healthy speakers and 6.0 s across the PD patients.

The signals are processed in a weighted overlap-add short-time Fourier transform (STFT) framework using a tight analysis window with a frame size of 48 ms (i.e., 768 samples) and an overlap of 50%.

B. Classifier Settings and Statistical Significance Analyses

For healthy and dysarthric speech classification, we use an SVM with a radial basis kernel function [7], [8], [16]. The validation strategy is a stratified 9-fold cross-validation, ensuring that each fold has the same number of healthy speakers and PD patients. In each fold, features are normalized using the mean and standard deviation of the training data. To select the soft margin constant C and the kernel width γ for the SVM, nested 5-fold cross-validation is performed on the training data in each fold with $C \in \{10^{-2},\ 10^{4}\}$, $\gamma \in \{10^{-4},\ 10^{2}\}$. The hyper-parameters used in each fold are selected as the ones resulting in the highest mean accuracy on the training data. Finally, the classification performance is evaluated in terms of the mean and standard deviation of the classification accuracy on the testing data across all folds.

To compare the performance of multiple SVMs (constructed using different feature vectors) in Sections VI-D, VI-E, and VI-F, statistical tests need to be conducted. Since training data across different folds overlap in a cross-validation framework, the test accuracy values obtained for an SVM in different folds are not independent. Consequently, usual statistical tests for multiple comparisons such as ANOVA [37] or the Friedman test [38] are inappropriate [39], [40]. To the best of our knowledge, a powerful statistical test applicable to comparing the performance of multiple classifiers in a cross-validation framework does not exist [39], [40]. In such a framework, only pairwise comparisons followed by multiple comparison corrections can be used [40]. To account for the lack of independence in the test accuracy values, we conduct pairwise comparisons using the corrected resampled t-test [41]. To control the family-wise error rate, Bonferroni-Holm corrections are applied [42]. Although such statistical analyses are presented in Sections VI-D, VI-E, and VI-F, it should be noted that these analyses should be taken with skepticism since their statistical power is yet to be determined.

C. Spectro-Temporal Sparsity of Healthy and Dysarthric Speech

In this section, the spectro-temporal sparsity of healthy and dysarthric speech spectral coefficients is compared.

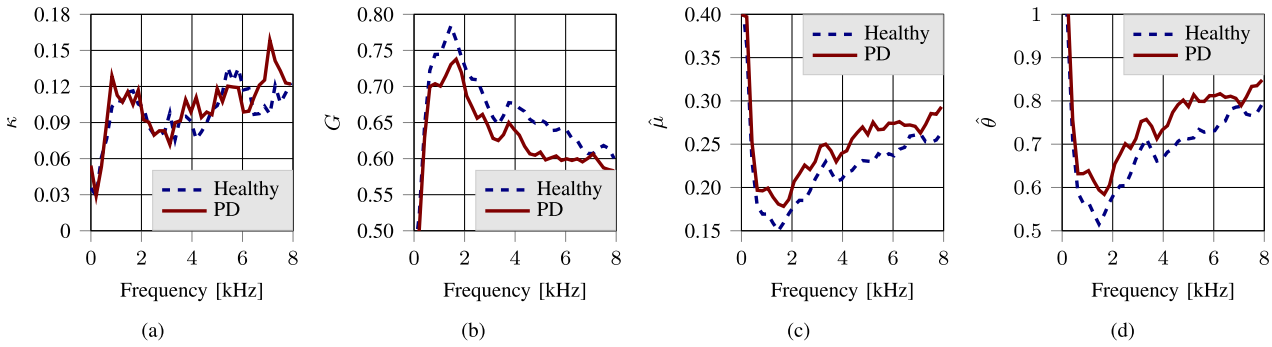

Fig. 7. Subband-dependent temporal sparsity estimate averaged across 45 healthy speakers and 45 PD patients using different sparsity measures: (a) kurtosis κ, (b) Gini index G, (c) estimated shape parameter of the Chi distribution $\hat{\mu}$, and (d) estimated shape parameter of the Weibull distribution $\hat{\theta}$. Sparsity measures for each speaker are estimated from a signal of 10 words, with a mean length of 6.1 s across the healthy speakers and 6.0 s across the PD patients.

The temporal sparsity vector is computed by concatenating all utterances for each speaker and estimating the sparsity in each subband. Since the STFT window size is 768 samples and taking into account only half of the spectrum, the temporal sparsity vector is an 385-dimensional vector. The computation of the spectral sparsity vector requires a reference speaker for alignment (cf. Section V). For the analyses presented in this section, we consider an additional (i.e., not part of the database described in Section VI-A) healthy male speaker r_1 as a reference speaker. The STFT representation of each utterance from all speakers in the database is first aligned to the STFT representation of the same utterance from r_1. For each speaker and each utterance, the spectral sparsity is estimated in all (aligned) time frames. Finally, for each speaker, the spectral sparsity estimates from all utterances are concatenated to create the spectral sparsity vector. Since the total number of time frames for all utterances from r_1 is 238, the spectral sparsity vector is an 238-dimensional vector (cf. Section V).

While the considered non-parametric sparsity measures can be directly computed (cf. Section III-A), numerical optimization routines on the speech spectral magnitudes should be used for the parametric sparsity measures (cf. Section III-B). For the results presented in the following, the optimization routines are initialized with $\mu = 1$ and $\theta = 1$.

Fig. 7 depicts the subband-dependent temporal sparsity estimates using kurtosis, Gini index, and shape parameters of the Chi and Weibull distributions. The measures are averaged across all healthy speakers and PD patients. Fig. 7(a) shows that the kurtosis, which is commonly investigated for dysarthric speech characterization, yields no distinct differences between the temporal sparsity of healthy and dysarthric speech. However, Fig. 7(b)–(d) show that the Gini index and the parametric sparsity measures yield clear differences between healthy and dysarthric speech across all subbands, with dysarthric speech being less temporally sparse than healthy speech.

To determine whether the previously discussed results are statistically significant, a statistical analysis is conducted. To compare the difference in mean temporal sparsity between healthy and dysarthric speech, an independent samples t-test is conducted for each subband. If the obtained p-value is smaller than 0.05, we consider there to be a statistically significant difference between the sparsity of healthy and dysarthric speech. Table I presents the percentage of subbands (out of 385) showing a statistically significant difference for each sparsity measure. As expected from the previous analysis, on the one hand the kurtosis measure results in the smallest number of subbands with a statistically significant difference between the temporal sparsity of healthy and dysarthric speech. On the other hand, the Gini index and the parametric sparsity measures show significant differences for many more subbands (ranging from 64.9% to 71.2%).

TABLE I
PERCENTAGE OF SUBBANDS (OUT OF 385) WHERE A STATISTICALLY SIGNIFICANT DIFFERENCE CAN BE FOUND BETWEEN THE TEMPORAL SPARSITY (CHARACTERIZED USING κ, G, $\hat{\mu}$, $\hat{\theta}$) OF HEALTHY AND DYSARTHRIC SPEECH. SIGNIFICANCE IS DETERMINED USING AN INDEPENDENT SAMPLES T-TEST AND THE STATISTICAL SIGNIFICANCE THRESHOLD IS CONSIDERED TO BE $p < 0.05$

κ	G	$\hat{\mu}$	$\hat{\theta}$
3.1%	64.9%	69.6%	71.2%

Fig. 8 depicts the time-dependent spectral sparsity estimates using kurtosis, Gini index, and shape parameters of the Chi and Weibull distributions. For visual convenience we present the measures for a single (arbitrarily selected) utterance, with the results generalizing to all other utterances. The measures are averaged across all healthy speakers and PD patients. Fig. 8(a) shows that when using kurtosis for characterizing the spectral sparsity of the considered utterance, dysarthric speech appears to be more sparse than healthy speech in all time frames. However, Fig. 8(b)–(d) show that when using the Gini index or the parametric sparsity measures, dysarthric speech appears to be more sparse than healthy speech only in some time frames (e.g., from $l = 1$ to $l = 3$). In other time frames (e.g., from $l = 8$ to $l = 10$), healthy speech appears to be more spectrally sparse than dysarthric speech. Informally inspecting and comparing the spectrograms of the considered utterance for several speakers suggests that from $l = 1$ to $l = 3$, dysarthric speech is typically dominated by articulation deficiencies, whereas from $l = 8$ to $l = 10$, dysarthric speech is typically dominated by breathiness and vocal tremor. On the one hand, articulation

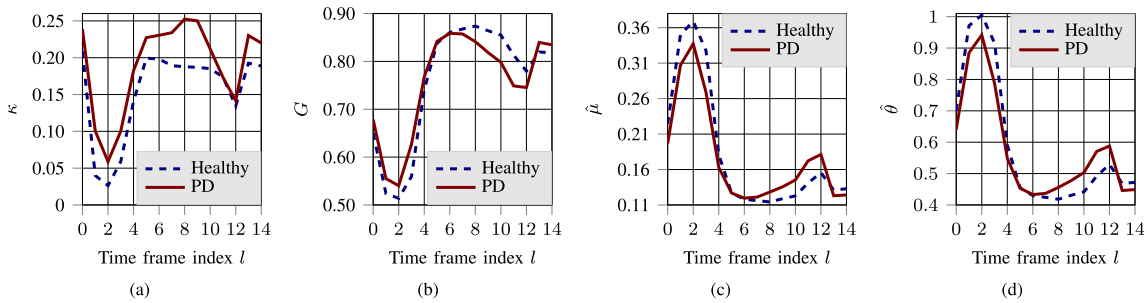

Fig. 8. Time-dependent spectral sparsity estimate for an exemplary utterance averaged across 45 healthy speakers and 45 PD patients using different sparsity measures: (a) kurtosis κ, (b) Gini index G, (c) estimated shape parameter of the Chi distribution $\hat{\mu}$, and (d) estimated shape parameter of the Weibull distribution $\hat{\theta}$.

TABLE II
PERCENTAGE OF TIME FRAMES (OUT OF 238) WHERE A STATISTICALLY SIGNIFICANT DIFFERENCE CAN BE FOUND BETWEEN THE SPECTRAL SPARSITY (CHARACTERIZED USING κ, G, $\hat{\mu}$, $\hat{\theta}$) OF HEALTHY AND DYSARTHRIC SPEECH. SIGNIFICANCE IS DETERMINED USING AN INDEPENDENT SAMPLES T-TEST AND THE STATISTICAL SIGNIFICANCE THRESHOLD IS CONSIDERED TO BE $p < 0.05$

κ	G	$\hat{\mu}$	$\hat{\theta}$
14.3%	26.1%	24.4%	26.1%

TABLE III
CLASSIFICATION ACCURACY [%] USING AN SVM WITH SPECTRAL SPARSITY ESTIMATES (CHARACTERIZED USING κ, G, $\hat{\mu}$, $\hat{\theta}$) FOR 5 DIFFERENT REFERENCE SPEAKERS

Sparsity measure	r_1	r_2	r_3	r_4	r_5	Mean $\pm$ std
κ	64.4	63.3	62.2	71.1	64.4	65.1 ± 3.1
G	75.6	74.4	75.6	81.1	72.2	75.8 ± 2.9
$\hat{\mu}$	76.7	66.7	74.4	76.7	71.1	73.1 ± 3.8
$\hat{\theta}$	73.3	68.8	80.0	75.6	70.0	73.6 ± 4.0

deficiencies can manifest as unexcited frequency components, resulting in dysarthric speech appearing to be more spectrally sparse than healthy speech. On the other hand, breathiness and vocal tremor can manifest as smearing of energy, resulting in dysarthric speech appearing to be less spectrally sparse than healthy speech. This evidence suggests that kurtosis yields an inaccurate characterization of the spectral sparsity of healthy and dysarthric speech, whereas the Gini index and the parametric sparsity measures yield a more accurate characterization.

To determine whether the previously discussed results are statistically significant, an independent samples t-test is conducted for each time frame. If the obtained p-value is smaller than 0.05, we consider there to be a statistically significant difference between the spectral sparsity of healthy and dysarthric speech. Table II presents the percentage of time frames (out of 238) showing a statistically significant difference for each sparsity measure. On the one hand, the kurtosis measure results in the smallest number of time frames with a statistically significant difference between the spectral sparsity of healthy and dysarthric speech. On the other hand, the Gini index and the parametric sparsity measures show significant differences for more time frames (ranging from 24.4% to 26.1%).

In summary, these results demonstrate that the spectro-temporal sparsity of dysarthric speech spectral coefficients differs from the spectro-temporal sparsity of healthy speech spectral coefficients. In addition, these results show that the Gini index and the parametric sparsity measures yield a more robust characterization of sparsity than the kurtosis measure.

D. Classifier Sensitivity to Reference Speaker Selection

Computing spectral sparsity requires a reference speaker to be selected beforehand for alignment (cf. Section V). Before investigating the advantages of using spectral sparsity as a feature vector for classifying healthy and dysarthric speech in Section VI-E, in this section we show that the performance of such a classifier is not sensitive to the used reference speaker.

We consider 5 healthy speakers r_i, $i = 1, \ldots, 5$, as reference speakers (with r_1 being the same reference speaker used in Section VI-C). These speakers are not part of the database described in Section VI-A, and hence, they do not appear in the training/testing data. For each r_i, the spectral sparsity feature vector for all speakers in the database is estimated based on the newly aligned STFT representations using kurtosis, Gini index, and parametric sparsity measures. The dimension of the spectral sparsity feature vector is 238 when using r_1, 234 when using r_2, 232 when using r_3, 278 when using r_4, and 294 when using r_5. The estimated spectral sparsity is used as a feature vector for an SVM as described in Section VI-B and the classification accuracy using different sparsity measures for different reference speakers is investigated.

Table III shows the classification accuracy when using the considered spectral sparsity measures for different reference speakers. In addition, the mean and standard deviation of the classification accuracy across all reference speakers are also presented. It can be observed that although different classification accuracies are obtained for different reference speakers, the performance is not highly sensitive to the reference speaker selection. The standard deviation of the classification accuracy for different reference speakers is low, ranging from 2.9% to 4.0%.

To determine whether a statistically significant difference exists between the performance of classifiers for different reference

TABLE IV
CORRECTED PAIRWISE COMPARISON RESULTS (p-VALUES) OF THE DIFFERENCE IN CLASSIFICATION ACCURACY OF SVMS WITH SPECTRAL SPARSITY ESTIMATES (CHARACTERIZED USING κ, G, $\hat{\mu}$, $\hat{\theta}$) FOR DIFFERENT REFERENCE SPEAKERS

	κ				G				$\hat{\mu}$				$\hat{\theta}$			
	r_2	r_3	r_4	r_5	r_2	r_3	r_4	r_5	r_2	r_3	r_4	r_5	r_2	r_3	r_4	r_5
r_1	0.937	0.937	0.303	0.999	0.929	0.999	0.468	0.447	0.324	0.760	0.999	0.530	0.434	0.343	0.663	0.588
r_2		0.937	0.119	0.937		0.929	0.156	0.852		0.530	0.362	0.749		0.333	0.333	0.851
r_3			0.119	0.937			0.245	0.727			0.760	0.749			0.333	0.211
r_4				0.300				0.156				0.396				0.333

TABLE V
MEAN AND STANDARD DEVIATION OF THE CLASSIFICATION ACCURACY [%] ACROSS ALL FOLDS USING AN SVM WITH TEMPORAL SPARSITY, SPECTRAL SPARSITY, AND SPECTRO-TEMPORAL SPARSITY ESTIMATES (CHARACTERIZED USING κ, G, $\hat{\mu}$, $\hat{\theta}$)

	κ	G	$\hat{\mu}$	$\hat{\theta}$
Temporal	52.2 ± 11.3	68.9 ± 13.7	65.6 ± 10.7	66.7 ± 9.4
Spectral	64.4 ± 6.8	75.6 ± 12.6	76.7 ± 12.5	73.3 ± 11.5
Spectro-temporal	64.4 ± 13.4	83.3 ± 6.7	80.0 ± 9.4	77.8 ± 9.2

speakers, the statistical analysis described in Section VI-B is conducted. For each sparsity measure, the corrected p-value for all pairwise comparisons of reference speakers is computed. Since 5 reference speakers are used, 10 pairwise comparisons are conducted for each measure. If the obtained p-value is smaller than 0.05, we consider there to be a statistically significant difference between the performance of the classifiers being compared.

Table IV shows the obtained p-value for each sparsity measure and all pairwise comparisons. It can be observed that all pairwise comparisons result in a p-value that is greater than 0.05. Hence, statistical tests suggest that the performance of an SVM using spectral sparsity is insensitive to the reference speaker selection, independently of the measure used to characterize spectral sparsity.

E. Classification Accuracy Using Spectro-Temporal Sparsity

In this section we compare the performance of an SVM aiming to discriminate between healthy and dysarthric speech using temporal sparsity, spectral sparsity, and spectro-temporal sparsity (characterized using different sparsity measures). The temporal sparsity feature vector is an 385-dimensional vector constructed as described in Section VI-C. The spectral sparsity feature vector is an 238-dimensional vector constructed as described in Section VI-D using the (arbitrarily selected) reference speaker r_1. The spectro-temporal sparsity feature vector is constructed by concatenating the temporal and spectral sparsity feature vectors, resulting in an 623-dimensional vector.

Table V presents the classification accuracy of an SVM using the different considered feature vectors. It can be observed that independently of the used sparsity measure, characterizing spectral sparsity yields a better classification accuracy than characterizing temporal sparsity. Furthermore, it can be observed that when using the Gini index and the parametric sparsity measures, characterizing both temporal and spectral sparsity improves the classification accuracy even further. As expected from the analyses in Section VI-C, Table V shows that the commonly used kurtosis measure (characterizing temporal, spectral, or spectro-temporal sparsity) yields the worst classification accuracy, whereas the Gini index and the parametric sparsity measures result in a significantly higher classification accuracy. When constructing the feature vector based on temporal sparsity, the highest classification accuracy of 68.9% is obtained using the Gini index. When constructing the feature vector based on spectral sparsity, the highest classification accuracy of 76.7% is obtained using the shape parameter of the Chi distribution. When constructing the feature vector based on spectro-temporal sparsity, the highest classification accuracy of 83.3% is obtained using the Gini index. Since computing the Gini index is computationally faster than the iterative ML estimation procedures required for the parametric sparsity measures, it can be said that the Gini index is an advantageous measure to use for spectro-temporal sparsity characterization.

It should be noted that using a nested cross-validation framework to find optimal hyper-parameters for each fold can yield positively biased final classification accuracy values. We have also analyzed the performance of the considered classifiers when hyper-parameters optimized for one fold are used for the complete database (i.e., 9 different hyper-parameter settings). While the final classification accuracy values can be slightly lower depending on the exact hyper-parameters, the standard deviation of the accuracy across different hyper-parameters is low. Most importantly, the performance trend across different features remains the same as in Table V.

To determine whether a statistically significant difference exists between the performance of SVMs using the different considered feature vectors, the statistical analysis described in Section VI-B is conducted. The corrected p-value for all pairwise comparisons is computed. Since 12 SVMs are used (i.e., 12 different sparsity-based feature vectors), 66 pairwise comparisons are conducted. If the obtained p-value is smaller than 0.05, we consider there to be a statistically significant difference between the performance of the SVMs being compared.

Table VI shows the obtained p-value for all pairwise comparisons, with bold entries indicating a significant difference at the considered threshold of 0.05. For notational convenience,

TABLE VI
CORRECTED PAIRWISE COMPARISON RESULTS (p-VALUES) OF THE DIFFERENCE IN CLASSIFICATION ACCURACY OF SVMS WITH DIFFERENT SPARSITY-BASED FEATURE VECTORS. BOLD ENTRIES INDICATE SIGNIFICANT DIFFERENCES AT A THRESHOLD OF 0.05

	κ_{S}	κ_{ST}	G_{T}	G_{S}	G_{ST}	$\hat{\mu}_{\mathrm{T}}$	$\hat{\mu}_{\mathrm{S}}$	$\hat{\mu}_{\mathrm{ST}}$	$\hat{\theta}_{\mathrm{T}}$	$\hat{\theta}_{\mathrm{S}}$	$\hat{\theta}_{\mathrm{ST}}$
κ_{T}	**0.030**	**0.081**	**0.019**	**0.003**	**0.001**	**0.019**	**0.001**	**0.001**	**0.013**	**0.001**	**0.001**
κ_{S}		0.999	0.472	**0.015**	**0.001**	0.852	0.062	**0.004**	0.604	0.150	**0.019**
κ_{ST}			0.671	**0.047**	**0.008**	0.899	0.177	0.081	0.768	0.212	0.124
G_{T}				0.518	0.081	0.398	0.449	0.209	0.544	0.679	0.361
G_{S}					0.150	0.333	0.840	0.472	0.290	0.604	0.746
G_{ST}						**0.009**	0.180	0.398	**0.011**	**0.001**	0.212
$\hat{\mu}_{\mathrm{T}}$							0.212	0.062	0.779	0.419	0.092
$\hat{\mu}_{\mathrm{S}}$								0.398	0.303	0.398	0.779
$\hat{\mu}_{\mathrm{ST}}$									0.092	0.180	0.449
$\hat{\theta}_{\mathrm{T}}$										0.449	0.180
$\hat{\theta}_{\mathrm{S}}$											0.427

subscripts are introduced for the different sparsity measures, with $\{\cdot\}_{\mathrm{T}}$, $\{\cdot\}_{\mathrm{S}}$, and $\{\cdot\}_{\mathrm{ST}}$ denoting the characterization of temporal sparsity, spectral sparsity, and spectro-temporal sparsity, respectively. Overall it can be observed that using kurtosis to characterize (temporal, spectral, or spectro-temporal) sparsity yields the most significant differences in classification accuracy when compared to using any other sparsity measure. More precisely, using κ_{T} yields a significantly worse classification accuracy than using any other measure. Further, using κ_{S} yields a significantly worse classification accuracy than using G_{S}, G_{ST}, $\hat{\mu}_{\mathrm{ST}}$, and $\hat{\theta}_{\mathrm{ST}}$. Finally, using κ_{ST} yields a significantly worse classification accuracy than using G_{S} and G_{ST}. The differences in classification accuracy when using the Gini index and the parametric sparsity measures to characterize (temporal, spectral, or spectro-temporal) sparsity often appear to be not significant, with the most significant differences arising when comparing G_{ST} to the remaining measures. Hence, it can be said that using G_{ST} yields a significantly better classification accuracy than using several other sparsity-based feature vectors such as $\hat{\mu}_{\mathrm{T}}$, $\hat{\theta}_{\mathrm{T}}$, or $\hat{\theta}_{\mathrm{S}}$.

In summary, the presented statistical analyses confirm the previously derived conclusions, i.e., characterizing sparsity using kurtosis yields the worst classification accuracy. Further, these analyses confirm that the Gini index and the parametric sparsity measures are more advantageous, with G_{ST} appearing to yield the most significant differences when compared to the remaining feature vectors. However, we advise the reader to treat these statistical analyses with caution since their statistical power is yet to be determined (cf. Section VI-B).

F. Classification Accuracy Using Spectro-Temporal Sparsity and State-of-the-Art Features

The results presented in Section VI-E show that using G_{ST} yields the highest classification accuracy among the proposed sparsity-based features. In this section, the classification accuracy obtained using G_{ST} is compared to using state-of-the-art features such as fundamental frequency f_0, jitter, shimmer, HNR, and MFCCs. We extract f_0, jitter, shimmer, HNR, and 15 MFCCs using the open-source toolkit openSMILE [43]. For each extracted quantity, feature vectors are constructed using 4 functionals, i.e., mean, standard deviation, kurtosis, and skewness [7], [16]. Hence, the feature vectors for f_0, jitter, shimmer, and HNR are 4-dimensional vectors, whereas the feature vector for MFCCs is an 60-dimensional vector (15 MFCCs $\times$ 4 functionals).

TABLE VII
MEAN AND STANDARD DEVIATION OF THE CLASSIFICATION ACCURACY [%] ACROSS ALL FOLDS USING AN SVM WITH THE PROPOSED G_{ST} FEATURE AND STATE-OF-THE-ART FEATURES

Feature	Performance
G_{ST}	83.3 ± 6.7
f_0	54.4 ± 13.4
jitter	52.2 ± 11.3
shimmer	57.8 ± 12.3
HNR	60.0 ± 11.5
MFCCs	76.7 ± 13.3

Table VII presents the classification accuracy of an SVM using G_{ST} and the considered state-of-the-art features.[2] It can be observed that out of all considered features, the proposed G_{ST} feature yields the highest mean classification accuracy of 83.3% and the lowest standard deviation of 6.7%. The classification accuracy obtained using f_0, jitter, shimmer, and HNR is generally low, ranging from 52.2% to 60.0%. Out of the considered state-of-the-art features, using MFCCs yields the highest classification accuracy of 76.7%, confirming the applicability of MFCCs to capture articulation deficiencies.

To determine whether a statistically significant difference exists between the performance of an SVM using G_{ST} and the

[2]The classification accuracy obtained using G_{ST} is clearly the same as in Section VI-E.

TABLE VIII
CORRECTED PAIRWISE COMPARISON RESULTS (p-VALUES) OF THE DIFFERENCE IN CLASSIFICATION ACCURACY OF SVMS WITH THE PROPOSED G_{ST} FEATURE AND STATE-OF-THE-ART FEATURES. BOLD ENTRIES INDICATE SIGNIFICANT DIFFERENCES AT A THRESHOLD OF 0.05

	f_0	jitter	shimmer	HNR	MFCCs
G_{ST}	**0.001**	**0.001**	**0.001**	**0.001**	0.179

considered state-of-the-art features, the statistical analysis described in Section VI-B is conducted. The corrected p-value for all pairwise comparisons is computed. Since 5 state-of-the-art feature vectors are considered, 5 pairwise comparisons are conducted. If the obtained p-value is smaller than 0.05, we consider there to be a statistically significant difference between the performance of SVMs using G_{ST} and the considered state-of-the-art feature.

Table VIII shows the obtained p-value for all pairwise comparisons, with bold entries indicating a significant difference at the considered threshold of 0.05. It can be observed that statistical analyses suggest that using G_{ST} yields a significantly better performance than using f_0, jitter, shimmer, or HNR. Further, these analyses suggest that the difference in classification accuracy when using G_{ST} and MFCCs is not significant. However, as described in Section VI-B, these statistical analyses need to be treated with skepticism.

VII. CONCLUSION

In this paper it has been proposed to use the spectro-temporal sparsity characterization as a robust feature for dysarthric speech detection. While characterizing the temporal sparsity is straightforward, to characterize the spectral sparsity it has been proposed to first align all available signals to the corresponding signals from an arbitrary selected reference speaker using DTW. The suitability of non-parametric sparsity measures (i.e., l_1-norm, kurtosis, Shannon entropy, and Gini index) and parametric sparsity measures (i.e., shape parameters of a Chi and Weibull distributions) for spectro-temporal sparsity characterization has been investigated. It has been shown that out of the considered measures, the Gini index and the parametric sparsity measures yield the most discriminative sparsity assessment of healthy and dysarthric speech. Further, it has been shown that compared to temporal sparsity, spectral sparsity is a more powerful discriminator for dysarthric speech detection.

ACKNOWLEDGMENT

The authors would like to thank the project partners from University of Paris III: Sorbonne Nouvelle, Geneva University Hospitals, and University of Geneva for a fruitful collaboration.

REFERENCES

[1] C. Marras *et al.*, "Prevalence of Parkinson's disease across North America," *Nature Partner J. Parkinson's Disease*, vol. 4, no. 21, Dec. 2018.
[2] J. R. Duffy, "Motor speech disorders: Clues to neurologic diagnosis," in *Parkinson's Disease and Movement Disorders: Diagnosis and Treatment Guidelines for the Practicing Physician*. Totowa, NJ, USA: Humana Press, 2000, pp. 35–53.
[3] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1015–1022, Apr. 2009.
[4] 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 Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, May 2012.
[5] D. Sztahó, G. Kiss, and K. Vicsi, "Estimating the severity of Parkinson's disease from speech using linear regression and database partitioning," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Dresden, Germany, Sep. 2015, pp. 498–502.
[6] 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. Annu. Conf. Int. Speech Commun. Assoc.*, San Francisco, CA, USA, Sep. 2016, pp. 1190–1194.
[7] J. R. Orozco-Arroyave *et al.*, "Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinson's disease," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Dresden, Germany, Sep. 2015, pp. 95–99.
[8] J. R. Orozco-Arroyave, F. Hönig, J. D. A.-L. no, J. F. Vargas-Bonilla, and E. Nöth, "Spectral and cepstral analyses for Parkinson's disease detection in Spanish vowels and words," *Expert Syst.: J. Knowl. Eng.*, vol. 32, no. 6, pp. 688–697, Dec. 2015.
[9] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, J. Villalba, J. R. Orozco-Arroyave, and N. Dehak, "Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's disease," *Appl. Soft Comput.*, vol. 62, pp. 649–666, Jan. 2018.
[10] S. Sapir, L. O. Ramig, J. L. Spielman, and C. Fox, "Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech," *J. Speech, Lang. Hearing Res.*, vol. 53, no. 1, pp. 114–125, Feb. 2010.
[11] S. Skodda, W. Visser, and U. Schlegel, "Vowel articulation in Parkinson's disease," *J. Voice*, vol. 25, no. 4, pp. 467–472, Jul. 2011.
[12] J. Rusz *et al.*, "Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task," *J. Acoustical Soc. Amer.*, vol. 134, no. 3, pp. 2171–2181, Sep. 2013.
[13] M. Novotný, J. Rusz, R. Čmejla, and E. Růžička, "Automatic evaluation of articulatory disorders in Parkinson's disease," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 9, pp. 1366–1378, Sep. 2014.
[14] J. R. Williamson *et al.*, "Segment-dependent dynamics in predicting Parkinson's disease," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Dresden, Germany, 2015, pp. 518–522.
[15] I. Kodrasi and H. Bourlard, "Statistical modeling of speech spectral coefficients in patients with Parkinson's disease," in *Proc. ITG Conf. Speech Commun.*, Oldenburg, Germany, Oct. 2018, pp. 271–275.
[16] I. Kodrasi and H. Bourlard, "Super-Gaussianity of speech spectral coefficients as a potential biomarker for dysarthric speech detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brighton, U.K., May 2019, pp. 6400–6404.
[17] Z. Xu and J. Sun, "Image inpainting by patch propagation using patch sparsity," *IEEE Trans. Image Process.*, vol. 19, no. 5, pp. 1153–1165, May 2010.
[18] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: From coding to source separation," *Proc. IEEE*, vol. 98, no. 6, pp. 995–1005, Jun. 2010.
[19] I. Kodrasi and S. Doclo, "Sparsity-promoting acoustic multi-channel equalization techniques," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 25, no. 7, pp. 1512–1525, Jul. 2017.
[20] H. Cheng, Z. Liu, L. Hou, and J. Yang, "Sparsity-induced similarity measure and its applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 4, pp. 613–626, Apr. 2016.
[21] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan, "Feature selection based on structured sparsity: A comprehensive study," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp. 1490–1507, Jul. 2017.
[22] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4723–4741, Oct. 2009.
[23] R. C. Hendriks, R. Heusdens, and J. Jensen, "Log-spectral magnitude MMSE estimators under super-Gaussian densities," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Brighton, U.K., Sep. 2009, pp. 1319–1322.
[24] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Proc. Int. Workshop Acoust., Echo, Noise Control*, Tel Aviv, Israel, Sep. 2010.
[25] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[26] R. Martin, "Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Orlando, FL, USA, May 2002, pp. 253–256.
[27] I. Tashev and A. Acero, "Statistical modeling of the speech signal," in *Proc. Int. Workshop Acoust., Echo, Noise Control*, Tel Aviv, Israel, Sep. 2010.
[28] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, Apr. 1995.
[29] R. Chartrand, "Shrinkage mappings and their induced penalty functions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, May 2014, pp. 1026–1029.
[30] A. Tkacenko and P. P. Vaidyanathan, "Generalized kurtosis and applications in blind equalization of MIMO channels," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2001, pp. 742–746.
[31] C. Fang, H. Li, L. Ma, and M. Zhang, "Intelligibility evaluation of pathological speech through multigranularity feature extraction and optimization," *Comput. Math. Methods Medicine*, vol. 2017, Jan. 2017, Art. no. 2431573.
[32] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans. Signal Process.*, vol. 47, no. 1, pp. 187–200, Jan. 1999.
[33] H. Dalton, "The measurement of the inequality of incomes," *Econ. J.*, vol. 30, no. 119, pp. 348–361, Sep. 1920.
[34] K. E. Atkinson, *An Introduction to Numerical Analysis*. Hoboken, NJ, USA: Wiley, 1989.
[35] J. R. Orozco-Arroyave, J. D. Arias-Londono, J. Vargas-Bonilla, M. Gonzalez-Rativa, and E. Noeth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in *Proc. Int. Conf. Lang. Resour. Eval.*, Reykjavik, Iceland, May 2014, pp. 342–347.
[36] G. T. Stebbins and C. G. Goetz, "Factor structure of the unified Parkinson's disease rating scale: Motor examination section," *Movement Disorders*, vol. 23, no. 7, pp. 633–636, Jul. 1998.
[37] R. A. Fisher, *Statistical Methods and Scientific Inference*. Edinburgh, U.K.: Oliver & Boyd, 1956.
[38] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *Ann. Math. Statist.*, vol. 11, no. 1, pp. 86–92, Mar. 1940.
[39] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006.
[40] G. Santafé, I. Inza, and J. A. Lozano, "Dealing with the evaluation of supervised classification algorithms," *Artif. Intell. Rev.*, vol. 44, pp. 467–508, Jun. 2015.
[41] C. Nadeau, and Y. Bengio, "Inference for the generalization error," *Mach. Learn.*, vol. 52, no. 3, pp. 239–281, Sep. 2003.
[42] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Statist.*, vol. 6, pp. 65–70, 1979.
[43] 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.

Ina Kodrasi (Member, IEEE) received the master of science degree in communications, systems, and electronics from Jacobs University Bremen, Bremen, Germany, in 2010, and the Ph.D. degree from the University of Oldenburg, Oldenburg, Germany, in 2015. She was a Research Associate and a Postdoctoral Researcher with the Signal Processing Group, University of Oldenburg from 2010 to 2015 and 2015 to 2017 respectively, where she worked on dereverberation and noise reduction. From 2010 to 2011, she was also with the Project Group Hearing, Speech and Audio Technology, Fraunhofer Institute for Digital Media Technology, Oldenburg, Germany, where she worked on microphone array beamforming. Since December 2017, she has been with the Idiap Research Institute, Martigny, Switzerland, working in the field of speech signal processing for clinical applications. She is member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and of the EURASIP Technical Area Committee on Acoustic, Speech and Music Signal Processing.

Hervé Bourlard (Fellow, IEEE) received the electrical and computer science engineering degree and the Ph.D. degree in applied sciences both from "Faculté Polytechnique de Mons," Mons, Belgium. Starting his research career as a member of the Scientific Staff with the Philips MBLE Research Laboratory of Brussels, he is currently (since 1996) Director of the Idiap Research Institute and a Full Professor with the Swiss Federal Institute of Technology Lausanne, Lausanne, Switzerland. He was also the Founding Director of the Swiss NSF National Centre of Competence in Research on "Interactive Multimodal Information Management (2001–2013)." Having spent (since 1988) several long-term and short-term visits at the International Computer Science Institute (ICSI), Berkeley, CA, he is currently an ICSI External Fellow and Emeritus Trustee.

He is the author/coauthor/editor of nine books and more than 350 reviewed (journal, conference, and book chapter) papers, including one IEEE Journal Paper Award. His main research interests include statistical pattern classification, signal processing, multi-channel processing, artificial neural networks, and applied mathematics, with applications to a wide range of information and communication technologies, including spoken language processing, speech and speaker recognition, language modeling, multimodal interaction, and augmented multi-party interaction.

He is an ISCA Fellow, a Senior Member of ACM, and an Elected Member of the Swiss Academy of Engineering Sciences. Having worked for academia as well as large and small (startup) industries, he is the recipient of several scientific and entrepreneurship awards.