Abstract: Epidemiological studies suggest that lifetime prevalence of voice disorders is about 30% for the general adult population. Moreover, vocal performance degradation may be amongst the earliest indicators of a neurodegenerative disease onset, such as Parkinson’s disease. Lacking alternative cost-effective biomarkers, biomedical speech signal processing has been gaining increasing impetus towards developing clinical decision support tools. Acoustic analysis of speech signals provides a convenient, automatic, accurate, robust, inexpensive, scalable approach assisting medical diagnosis and symptom severity monitoring. Nevertheless, the algorithmic tools developed for biomedical speech signal processing are spread across different software platforms, hindering direct algorithmic comparisons and the further development of this impending field. This study brings many biomedical speech signal processing algorithms together under the same software platform, and has led to the development of a practical free toolkit which can be accessed over the Internet using a simple application.

Keywords: Acoustic analysis, decision support tool, speech signal processing, sustained vowels

I. INTRODUCTION

Neurodegenerative disorders and speech performance degradation have been closely associated at least since the 1970’s. For example, the vast majority of Parkinson’s disease (PD) subjects experience some problems associated with their voice [1]. In addition, the lifetime prevalence of general voice disorders is approximately 30% for the adult population, and have a substantial impact on a person’s personal and professional life [2]. Regardless of pathological cause, voice disorders are characterized by the malfunction of one or more components in the vocal production mechanism, leading to poor vocal quality. Depending on the pathological cause, characteristic symptoms may include reduced or increased loudness, vocal tremor, and breathiness amongst others.

In speech clinical practice, human experts assess a subject’s voice quality using sustained vowel phonations, and/or conversational speech. Sustained vowels, where the subject is asked to prolong his phonation for as long as possible and as steady as possible (in terms of pitch and amplitude), are particularly practical because they circumvent linguistic artifacts and are considered sufficient for many voice assessment applications [3].

The lack of a sufficient number of experts to perform vocal assessments has prompted the development of biomedical speech signal processing algorithms, to objectively and automatically characterize clinically useful properties of the speech signals. There is considerable research on the topic of developing clinical decision support tools using speech signals, in particular to study

Fig. 1. Methodology for biomedical speech signal processing.
neurodegenerative disorders [4-11]. The nature of this multi-disciplinary domain has attracted the attention of speech experts, phoneticians, clinicians, mathematicians, and engineers, and the algorithmic tools are dispersed in the literature across research areas. In many cases there is no readily freely available implementation of the algorithms; alternatively, the source code is provided by its developers in diverse software platforms, hindering their widespread use by practitioners.

This study describes the framework for the development of an automated clinical decision support tool, bringing together many of the biomedical speech signal processing algorithms which were previously scattered in the research literature. It provides the conceptual basis for the development of the current state of the art biomedical speech signal processing algorithms, highlighting the intricate characteristics of the speech signals which the algorithmic tools aim to characterize. Finally, there are some suggestions about directions future work could take in the quest of extracting clinically useful information from the speech signals which may not be adequately quantified using the currently available algorithmic tools.

II. METHODS

The methodology of a clinical decision support tool is summarized in Fig. 1. It consists of processing the original raw time series (speech signal) to extract distinctive, clinically useful properties (feature calculation), selecting or transforming the computed speech signal properties (feature selection or feature transformation), and mapping the final compact feature subset to the clinical output we want to associate the speech signal with (feature mapping). It is critical to emphasize that the developed algorithms have only been validated in the sustained vowel /a/ setting; it is not clear how meaningful and useful the output of the presented algorithmic tools is for other settings.

This study focuses on the computation of the speech signal properties and very briefly outlines the other two components.

A. Computation of features

This section summarizes briefly the most widely used biomedical speech signal processing algorithms, clustering them into algorithmically-related families. For a detailed overview, see Tsanas [11]. In principle, any signal processing tool and any time series analysis tool could be used to extract characteristics from the speech signal which might be useful for clinical applications.

Many speech signal processing algorithms rely on the accurate computation of the fundamental frequency (F0) which is critical to characterize speech signals [12]. Defining F0 is not straightforward in non-periodic signals (i.e. all pathological cases) and the development of robust F0 estimators is intensively pursued [13]. On the current evidence, the NDF [14] and SWIPE F0 estimators [15] appear very promising in processing sustained /a/ vowels, based on extensive comparisons against reference synthetic speech data created by a state of the art physiologica model [11].

The first algorithmic group builds on physiological evidence that the vocal folds’ oscillating pattern is nearly periodic in healthy voices and substantially departs from periodicity in pathological cases [3]. Two of the most well-known algorithms, jitter and shimmer, belong to this group [16], [3]. Jitter quantifies F0 deviations, and shimmer quantifies amplitude deviations. There is no unique formal definition of jitter and shimmer, and researchers have developed many jitter variants and shimmer variants. Similarly, the Recurrence Period Density Entropy (RPDE), the Pitch Period Entropy (PPE), the Glottal Quotient (GQ), and other F0-related measures [11] build on the concept of quantifying the extent of aperiodicity in the vocal folds’ oscillating pattern.

The second general algorithmic group comprises signal to noise ratio (SNR) approaches. The rationale is that incomplete vocal fold closure leads to the creation of aerodynamic vortices, leading to increased acoustic noise. Harmonic to Noise Ratio (HNR), Detrended Fluctuation Analysis (DFA), Glottal to Noise Excitation (GNE), Vocal Fold Excitation Ratio (VFER), and Empirical Mode Decomposition Excitation Ratio (EMD-ER) express this concept algorithmically.

Linear Predicting Coding Coefficients (LPCC) is a generic signal processing tool: the underlying concept is that pathological voices are be more irregular and hence more difficult to predict based on past samples. Wavelets are another generic tool to analyzing non-stationary time series signals: we proposed using wavelet decomposition to analyze the F0 contour, and use the wavelet coefficients as features [17]. Lastly, Mel Frequency Cepstral Coefficients (MFCCs) are the gold standard in speaker recognition, and have recently shown great promise in biomedical applications [18], [9], [10].

The key information of the biomedical speech signal processing algorithms included in the toolkit is summarized in Table 1.

B. Feature selection or feature transformation

A common problem in applications with many features is the curse of dimensionality: a compact feature subset may lead to improved performance and promotes interpretability by means of inferring the most prominent properties of the speech signals for the investigated problem [19]. There are two main approaches: feature selection (selecting a subset of the original features), and feature transformation (transforming the original features to develop new, more predictive features). This falls outside the scope of this study; we refer to Guyon et al. [20] and Hastie et al. [19] for a comprehensive and authoritative overview.
Table 1: Summary and key information of the biomedical speech signal processing algorithms in the toolkit

<table>
<thead>
<tr>
<th>Measure</th>
<th>Motivation</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jitter &amp; Jitter variants</td>
<td>The vocal folds are affected in voice disorders, and jitter aims to capture instabilities of the oscillating pattern of the vocal folds quantifying the cycle-to-cycle changes in fundamental frequency</td>
<td>One for each variant</td>
</tr>
<tr>
<td>Shimmer &amp; shimmer variants</td>
<td>Shimmer aims to capture instabilities of the oscillating pattern of the vocal folds quantifying the cycle-to-cycle changes in amplitude</td>
<td>One for each variant</td>
</tr>
<tr>
<td>Recurrence Period Density Entropy (RPDE)</td>
<td>Quantifies the stochastic component of the deviation of vocal fold periodicity</td>
<td>1</td>
</tr>
<tr>
<td>Pitch Period Entropy (PPE)</td>
<td>In speech disorders it is very difficult to sustain stable pitch due to incomplete vocal fold closure. PPE quantifies the impaired control of stabilised pitch.</td>
<td>1</td>
</tr>
<tr>
<td>$F_0$-related measures</td>
<td>Summary statistics of $F_0$, differences from expected $F_0$ in age- and gender-matched controls, variations in $F_0$</td>
<td>Three for each $F_0$ estimation algorithm</td>
</tr>
<tr>
<td>Harmonics to Noise Ratio (HNR) &amp; Noise to Harmonics Ratio (NHR)</td>
<td>In speech pathologies there is increased noise due to turbulent airflow, resulting from incomplete vocal fold closure. HNR and NHR quantify the ratio of actual signal information over noise.</td>
<td>4</td>
</tr>
<tr>
<td>Detrended Fluctuation Analysis (DFA)</td>
<td>Quantifies the stochastic self-similarity of the noise caused by turbulent airflow</td>
<td>1</td>
</tr>
<tr>
<td>Glottal to noise excitation (GNE)</td>
<td>Extent of noise in speech using energy and nonlinear energy concepts</td>
<td>6</td>
</tr>
<tr>
<td>Vocal fold excitation ratio (VFER)</td>
<td>Extent of noise in speech using energy, nonlinear energy, and entropy concepts</td>
<td>9</td>
</tr>
<tr>
<td>Empirical mode decomposition excitation ratio (EMD-ER)</td>
<td>Signal to noise ratios using EMD-based energy, nonlinear energy and entropy</td>
<td>6</td>
</tr>
<tr>
<td>Linear Predicting Coding Coefficients (LPCC)</td>
<td>Quantify deviations of the prediction of the current data sample as a function of the preceding samples. In pathological voices this deviation is expected to be larger.</td>
<td>10</td>
</tr>
<tr>
<td>Wavelet measures</td>
<td>Quantify deviations in $F_0$ (obtained using any $F_0$ estimation algorithm)</td>
<td>180</td>
</tr>
<tr>
<td>Mel Frequency Cepstral Coefficients (MFCC)</td>
<td>Voice pathologies lead to decreased control of the articulators (vocal tract), and the MFCCs attempt to analyse the vocal tract independently of the vocal folds</td>
<td>12-42, depends on extracted components and the use of delta and delta-delta coefficients</td>
</tr>
</tbody>
</table>

C. Feature mapping

So far, we have described the methodology towards extracting speech signal properties (features), and determining a robust representation of the extracted information. In order to develop a complete decision support tool it is necessary to associate the computed features with the clinical outcome of interest. In order to achieve this, a database of labeled data is required: the user has to provide a database of speech signals where each speech signal (the *.wav file) corresponds to the known clinical outcome. Then, a supervised learning mapping algorithm associates the features with the clinical outcome (response). The aim is to subsequently use unlabeled data and interrogate the mapping algorithm to provide estimates of the labels. One suggestion is using Random Forests, a robust statistical machine learning tool which has shown impressive performance without requiring any hyper-parameter optimization. We refer to Hastie et al. [19] for details.

III. RESULTS

The toolkit requires a single input (*.wav file) and provides an output vector with the computed features along with the corresponding feature names. Optionally, it can also output the F0 contour for visualization and further processing.

IV. DISCUSSION

This study brings together many biomedical speech signal processing algorithms under a common software platform, resulting in the development of a new toolkit. This may facilitate further advances in extracting clinical-
ly useful information from speech signals, enabling practitioners easily apply known algorithms to their data, and researchers compare new algorithmic concepts to the existing literature. Hitherto, the existing speech signal processing algorithms were dispersed in the literature across research areas, a fact which reflects the multidisciplinary nature of this field. Although some researchers have open-sourced their algorithmic contributions, there was no standardized software platform to use these tools so far.

The toolkit has only been validated on sustained vowel /a/ phonations; its functionality may be extended to additional vowels and their interaction, for example to compute the vowel space area and related algorithms. The default settings have been optimized for biomedical applications in Parkinson’s disease [9-11], but the toolkit is flexible to allow experienced users adjust additional parameters. Currently, there is no provision to extend the analysis to conversational speech, although this might be an interesting approach to pursue in the future.

A critical aspect of biomedical speech signal processing is interpretability. That is, determining the most distinctive properties of the speech signals for the investigated application, be means of observing the (jointly) most predictive subset of speech signal processing algorithms. To this end, it is often advisable to investigate many diverse feature selection algorithms to gain insight into the pathophysiological voice characteristics which are mostly predictive of the investigated clinical outcome [21]. Moreover, given a sufficiently large sample size, it is often useful to stratify the data by gender [9], [22].

Future studies could further investigate applying robust time-series analysis tools to this domain. However, probably the most promising approach towards developing new biomedical speech signal processing tools may be studying closely the physiology of the investigated application. Different voice disorders have different pathophysiological characteristics, and improved acoustic results might be obtained by developing new physiologically-informed algorithms targeting specific characteristics.

REFERENCES