Quality Control of Voice Recordings in Remote Parkinson's Disease Monitoring using the Infinite Hidden Markov Model

Alavijeh, Amir Hossein Poorjam; Raykov, Yordan P.; Badawy, Reham; Jensen, Jesper Rindom; Christensen, Mads Græsbøll; Little, Max A.

Published in:
2019 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2019 - Proceedings

DOI (link to publication from Publisher):
10.1109/ICASSP.2019.8682523

Publication date:
2019

Document Version
Accepted author manuscript, peer reviewed version

QUALITY CONTROL OF VOICE RECORDINGS IN REMOTE PARKINSON'S DISEASE MONITORING USING THE INFINITE HIDDEN MARKOV MODEL

Amir Hossein Poorjam [1]*, Yordan P. Raykov* [2]*, Reham Badawy* [2]*,*
Jesper Rindom Jensen [1]*, Mads Græsbøll Christensen* [1] *and Max A. Little* [2,3]

[1] Audio Analysis Lab, CREATE, Aalborg University, Aalborg, DK
[2] Engineering and Applied Sciences, Aston University, Birmingham, UK
[3] Media Lab, MIT, Cambridge, Massachusetts, USA

[1] {ahp,jrj,mgc}@create.aau.dk, [2] yordan.raykov@gmail.com, [2] rehambadawy@hotmail.com,
[2,3] max.little@aston.ac.uk

ABSTRACT

The performance of voice-based systems for remote monitoring of Parkinson's disease is highly dependent on the degree of adherence of the recordings to the test protocols, which probe for specific symptoms. Identifying segments of the signal that adhere to the protocol assumptions is typically performed manually by experts. This process is costly, time consuming, and often infeasible for large-scale data sets. In this paper, we propose a method to automatically identify the segments of signals that violate the test protocol with a high accuracy. In our approach, the signal is first split into variable duration segments by fitting an infinite hidden Markov model (iHMM) to the frames of the signals in the mel-frequency cepstral domain. The complexity of the iHMM is capable of growing jointly with the data allowing us to infer a potentially large (asymptotically infinite) number of different phenomena segmented into different hidden states. Then, we identify the segments that adhere to the test protocol by applying a multinomial naive Bayes classifier to the state indicators of segments. The experimental results show that even by using a small amount of training data, we can achieve around 96% accuracy in identifying short-term protocol violations with a 0.2 s resolution.

Index Terms— Bayesian Nonparametric, infinite HMM, Parkinson's disease, quality control, segmentation

1. INTRODUCTION

Advances in speech signal analysis and machine learning facilitate the development of highly accurate, data-driven techniques for detecting Parkinson's disease (PD) symptoms from speech signals [1–4]. The development of a reliable PD detection system requires large amounts of training data to accurately model the characteristics of the speech signals that are indicative of PD. The performance of such systems is highly affected by the quality of the recordings [3,4]. Collecting high quality voice samples typically requires participants to be present in clinic to record their voice in controlled experimental conditions. However, creating large data sets this way is challenging and is infeasible in practice where behavioural and environmental confounding factors may infiltrate the recordings. Moreover, monitoring the progression of PD requires voice samples recorded regularly from each subject which makes this controlled data collection impractical in long term.

This work was funded by Independent Research Fund Denmark: DFF 4184-00056

Remote monitoring of PD symptoms bypasses the logistical limitations of controlled data collection in clinic, in which participants are asked to follow a set of instructions that probe for specific voice symptoms via a web-based interface, or an application running on a portable device such as a smartphone [5, 6]. Moreover remote monitoring provides participants the flexibility to record their voice at any time and location, resulting in a larger population sample. However, outside controlled lab conditions, there is a higher risk that participants may violate the test protocol during recording due to lack of training, misinterpretation of the test protocol or negligence. Processing segments of the recordings which do not comply with the assumptions of the test protocol can produce misleading, non-replicable and non-reproducible results [7, 8] that could have significant ramifications for the patients' health. Therefore, it is pivotal to be able to identify parts of the signal in which the patient is adhering to the test protocol. Those parts of the recording that violate the test protocol may be either excluded from the data analysis or be enhanced [9]. In clinic, quality control of voice recordings is typically performed manually by human experts. This process, however, is very costly, time consuming, and often infeasible for large-scale data sets. Thus, there is a need for automatic quality control approaches in remote voice monitoring of PD symptoms.

Several attempts have been made to address quality control on pathological voices. The simplest approach, which is widely used in speech applications, is the voice activity detection (VAD) in which the voiced and unvoiced parts of a speech signal are identified [10]. However, the performance of many VAD systems is adversely influenced by increasing the background noise level or by the presence of unexpected degradations [10, 11]. In [12, 13], the problem of quality control in pathological voices has been approached as a classification task in which different types of degradation that are commonly present during recording or transmission are classified. However, the performance of these approaches is limited when new degradation types are introduced. More importantly, protocol violations are not limited to the presence of degradations in signals. For example, talking, laughing or coughing may all be considered as protocol violations in sustained vowel data sets even though the recordings are of high quality. Badawy et al. [14] proposed a general framework for detecting a wide range of protocol violations using a nonparametric switching autoregressive (AR) model; however, low order AR models fail to capture variations in the low frequency harmonics.

In contrast, we have developed a framework in this paper which fits an infinite hidden Markov model (iHMM) to the frames of the

recordings in the mel-frequency cepstral domain, resulting in splitting the voice recording into segments of variable duration in an unsupervised manner. Unlike the parametric HMM, the number of states in the iHMM is not fixed a priori, and thus our model can automatically adapt to the complexity of the data. This facilitates discovering a variety of events in the voice recordings. A simple multinomial naive Bayes classifier is then applied to identify which segments are associated with protocol adherence or violation.

2. SYSTEM DESCRIPTION

2.1. Problem Formulation

Automatic quality control of remotely collected pathological voice recordings can be viewed as the identifying segments of the voice recordings that adhere to the test protocol. An example of such protocols for remote monitoring of PD symptoms using smartphones is as follows: participants are required to hold the phone in a similar position to making a phone call, take a deep breath and utter a sustained vowel /a/ at a comfortable, steady pitch and intensity for as long as they can in a low-noise environment. Thus, other events than a sustained vowel phonation such as user interactions with the smartphone during the voice test, recording in a noisy environment and irrelevant activities such as talking, laughing and coughing are all considered to be protocol violations [5, 6].

To control the protocol adherence or violation in PD voice data sets, we are given an ordered set of tuples $\mathcal{X} = \big((\boldsymbol{x}_t, y_t)\big)_{t=1}^{T}$ as the training data where $\boldsymbol{x}_t \in \mathbb{R}^d$ is the t^{th} observation of d dimension and $y_t \in \{1, 2\}$ denotes the corresponding adherence/violation label. The goal is to estimate a classifier function such that for an observation not in the training data, the probability of the estimated output being classified to the correct class is maximized.

2.2. Segmentation with the Infinite Hidden Markov Model

HMMs are widely used for modelling time-dependent patterns such as speech and language [15]. A HMM represents a probability distribution over sequences of observations $\boldsymbol{x}_{1:T} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T)$ of length T by invoking a Markov chain of hidden state variables $s_{1:T} = (s_1, \ldots, s_t, \ldots, s_T)$ where each s_t is in one of the K possible states [16]. In the first-order HMM, the distribution of the state at time t depends on the state immediately before it, and the transition conditional is parameterized by a $K \times K$ transition matrix $\boldsymbol{\pi}$ where $\pi_{ij} = P(s_t = j | s_{t-1} = i)$, for $i, j = 1, 2, \ldots, K$. The likelihood of the observation $\boldsymbol{x}_t$ is modeled with a distribution of K mixture components as:

$$P(\boldsymbol{x}_t | s_{t-1} = i, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_{i,k} P(\boldsymbol{x}_t | \boldsymbol{\theta}_k), \tag{1}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ are the time-independent emission parameters which parameterize the observation model f for each state; that is, $\boldsymbol{x}_t | s_t \sim f(\boldsymbol{\theta}_{s_t})$ is a draw from a distribution $f(\boldsymbol{\theta}_{s_t})$. If we assume that similar phenomena in recordings of a voice data set have the tendency to be clustered together, by identifying the HMM states with the clusters, one can use a HMM to cluster the observations in terms of different events. However, the key problem here is that we do not have prior knowledge about the number of events (i.e. states) that can be present in the recordings. Although there exist a variety of techniques for choosing the number of states [17–19], it is still challenging to predict how many states are required to cover all events such that we do not encounter unobserved events in the future. Furthermore it is reasonable to assume that as we observe more data, different types of protocol violations will appear and thus the inherent number of states will have to adapt accordingly. Therefore, we use a nonparametric Bayesian approach to relax the assumption of a fixed K in (1). The simplest approach is to use a symmetric Dirichlet prior with parameter α/K over the transition probabilities and take $K \to \infty$ [20]. However, since the transition probabilities will have independent priors, there will be no coupling across transitions from different states [21]. To tackle this problem, we can use hierarchical Dirichlet priors which have shared parameters:

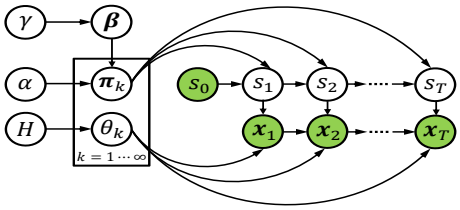

Fig. 1: The graphical model for the infinite HMM

$$\boldsymbol{\beta} \sim \text{Dirichlet}\left(\frac{\gamma}{K}, \cdots, \frac{\gamma}{K}\right), \quad \boldsymbol{\pi}_k \sim \text{Dirichlet}(\alpha\boldsymbol{\beta}), \tag{2}$$

where $\boldsymbol{\beta}$ are the shared prior parameters and $\boldsymbol{\pi}_k$ are transition probabilities from state k. Teh et al. showed that as $K \to \infty$, the hierarchical prior in (2) becomes a hierarchical Dirichlet process which is a set of Dirichlet processes (DPs), $G_k \sim \text{DP}(\alpha, G_0)$, with a local concentration parameter $\alpha > 0$, that are linked together with a shared random base measure, $G_0 \sim \text{DP}(\gamma, H)$, drawn from a DP with a global concentration parameter $\gamma > 0$ and a global base measure H [22]. H is the global base distribution over the component parameters of the HMM. α and γ can be viewed as prior counts for the local and global DPs, respectively. These random measures can, under the stick-breaking representation [23], be formulated as:

$$G_0 = \sum_{j=1}^{\infty} \beta_j \delta_{\boldsymbol{\theta}_j}, \quad G_k = \sum_{j=1}^{\infty} \pi_{kj} \delta_{\boldsymbol{\theta}_j}, \quad (k = 1, 2, \ldots, \infty), \tag{3}$$

where $\boldsymbol{\pi}_k \sim \text{DP}(\alpha, \boldsymbol{\beta})$, each $\boldsymbol{\theta}_j$ is a sample drawn independently from H, $\delta_{\boldsymbol{\theta}_j}$ denotes an atom at $\boldsymbol{\theta}_j$, and $\boldsymbol{\beta} = (\beta_j)_{j=1}^{\infty} \sim \text{GEM}(\gamma)$ is the stick-breaking representation for DPs which is drawn from Griffiths-Engen-McCloskey distribution with parameter γ [23]. This implies that there is a different measure G_k related to each row of the transition matrix which associates different weights $\pi_{k,1}, \ldots, \pi_{k,K}$ in the transition matrix. It also indicates that for discrete G_0 (which it has to be when $G_0 \sim \text{DP}$), the G_k's share the component parameters $\boldsymbol{\theta}_k$ and a posteriori each G_k has finite support at a subset of $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$. This brings us to the point where we can define the infinite hidden Markov model (iHMM), which can possibly have countably infinite number of hidden states, as follows:

$$\begin{aligned}
\boldsymbol{\beta} &\sim \text{GEM}(\gamma) \\
\boldsymbol{\pi}_k &\sim \text{DP}(\alpha, \boldsymbol{\beta}) && (k = 1, 2, \ldots, \infty) \\
\boldsymbol{\theta}_k &\sim H && (k = 1, 2, \ldots, \infty) \\
s_0 &= 1 \\
s_t | s_{t-1} &\sim \boldsymbol{\pi}_{s_{t-1}} && (t = 1, 2, \ldots, T) \\
\boldsymbol{x}_t | s_t &\sim f(\boldsymbol{\theta}_{s_t}) && (t = 1, 2, \ldots, T).
\end{aligned} \tag{4}$$

The graphical model for the iHMM defined in (4) is shown in Fig. 1. The hyper-parameters α and γ play an important role in controlling the number of states in the iHMM. Particularly, γ controls the probability of generating new states; that is, choosing a large value for γ results in producing a large number of states. On the

other hand, α controls the sparsity of the transition matrix by controlling the transitions between different states; that is, setting α to a small number leads to increasing the probability of taking the existing transitions.

We use the direct assignment Gibbs sampler for inferring the posterior over the sequence of hidden states [22]. The Gibbs sampler marginalizes out the infinitely many transition parameters $\boldsymbol{\pi}$ and emission parameters $\boldsymbol{\Theta}$. In each iteration of the Gibbs sampling, we first re-sample the hidden states and then the base distribution parameters. To re-sample the sequence of $s_{1:T}$, we take out one s_t at a time and re-sample it from the posterior

$$P(s_t|s_{t'}, \boldsymbol{x}_{1:T}, \alpha, \gamma, \boldsymbol{\beta}, H, f) \propto P(\boldsymbol{x}_t|s_t, s_{t'}, \boldsymbol{x}_{t'}, H, f) \\ P(s_t|s_{t'}, \alpha, \gamma, \boldsymbol{\beta}), \tag{5}$$

where $\boldsymbol{x}_{t'}$ denotes all observations except $\boldsymbol{x}_t$, and $s_{t'}$ indicates all states except s_t. If f is conjugate to H, there is a closed-form solution to calculate the first term on the right-hand side of (5), which is the conditional likelihood of $\boldsymbol{x}_t$:

$$P(\boldsymbol{x}_t|s_t, s_{t'}, \boldsymbol{x}_{t'}, H, f) = \int P(\boldsymbol{x}_t|\boldsymbol{\theta}_{s_t}) P(\boldsymbol{\Theta}|s_{t'}, \boldsymbol{x}_{t'}, H) d\boldsymbol{\Theta}. \tag{6}$$

As the hidden state sequence is Markov, we can calculate the second term on the right-hand side of (5) as:

$$P(s_t = k|s_{t'}, \alpha, \boldsymbol{\beta}) \propto \\ \begin{cases} (n_{s_{t-1},k} + \alpha\beta_k)\frac{n_{k,s_{t+1}} + \alpha\beta_{s_{t+1}}}{n_{k,\cdot} + \alpha} & \text{for } k \leq K, s_{t-1} \neq k \\ (n_{s_{t-1},k} + \alpha\beta_k)\frac{1 + n_{k,s_{t+1}} + \alpha\beta_{s_{t+1}}}{1 + n_{k,\cdot} + \alpha} & \text{for } s_{t-1} = s_{t+1} = k \\ (n_{s_{t-1},k} + \alpha\beta_k)\frac{n_{k,s_{t+1}} + \alpha\beta_{s_{t+1}}}{1 + n_{k,\cdot} + \alpha} & \text{for } s_{t-1} = k \neq s_{t+1} \\ \alpha\beta_k\beta_{s_{t+1}} & \text{for } k = K+1, \end{cases} \tag{7}$$

where $n_{i,j}$ denotes the number of transitions from state i to state j excluding the time steps $t-1$ and t, $n_{i,\cdot}$ stands for the total transitions from state i, and K is the number of states in $s_{t'}$.

According to (3), $\boldsymbol{\beta}$ contains the mixture weights. If we combine the weights of all unrepresented components $\beta_{K+1}, \ldots, \beta_\infty$ into the term $\beta_{K+1} = \sum_{k=K+1}^{\infty} \beta_k$, then $\boldsymbol{\beta}$ can be re-sampled from the posterior $(\beta_1, \ldots, \beta_{K+1}) \sim \text{Dirichlet}(m_1, \ldots, m_K, \gamma)$, where m_k denotes the number of times the transition to state k has been drawn from the global DP. For more details about the Gibbs sampling, we refer to [22].

We propose to fit the iHMM to the mel-frequency cepstral coefficients (MFCCs) extracted from short time frames of voice signals. The motivation for using cepstral features is that not only do they convey information about speech content [24], but we have also shown in [12,25] that degradation in speech signals predictably modifies the distribution of the MFCCs by changing the covariance of the features and shifting the mean to different regions in feature space.

2.3. Classification of the Hidden States

The states of the iHMM correspond to different events in the recordings. Thus, segments of the voice recordings with similar characteristics are clustered together under the same state indicator values. This facilitates a better understanding of the changes in signal characteristics due to, for example, the presence of different types of signal degradation, speaker variability, protocol violations, and vocal disorders. In this paper, we are interested in identifying the segments of the signal that are sufficiently reliable for detecting PD voice symptoms. This can be performed by detecting the states which adhere to and those that violate the voice test protocols described in Section 2.1. We propose to use the multinomial naive Bayes (MNB) classifier to map the state indicators $s_{1:T}$ to the binary labels $y_{1:T} = (y_1, \ldots, y_t, \ldots, y_T)$, where $y_t = 1$ if $\boldsymbol{x}_t$ complies with the protocol or $y_t = 2$ if it violates the protocol. The multinomial naive Bayes is a probabilistic classifier which assumes that the data points in different classes have different multinomial distributions. In the MNB classifier, a feature vector for the t^{th} observation $\boldsymbol{\rho}_t = (\rho_{t,1}, \ldots, \rho_{t,K})$ is a histogram, with $\rho_{t,k}$ being the number of times state k is observed. Assuming that the most probable event at time t is modeled by state k, we can use the modal estimates of the state indicators as the features along with the corresponding binary labels to train the classifier. The likelihood of the histogram of a new observation $\tilde{\boldsymbol{\rho}}$ is defined as:

$$P(\tilde{\boldsymbol{\rho}}|y_{1:T}, \tilde{y}, \boldsymbol{\rho}_{1:T}) = \frac{(\sum_{k=1}^{K} \rho_{t,k})!}{\prod_{k=1}^{K} \rho_{t,k}!} \prod_{k=1}^{K} p_{k,\tilde{y}}^{\rho_{t,k}}, \tag{8}$$

where $p_{k,\tilde{y}}$ is the probability of the k^{th} attribute being in class $\tilde{y} \in \{1, 2\}$, which is trained using the training data. Using the Bayes rule and the prior class probability $P(\tilde{y})$, the class label for a new test observation is predicted as:

$$\hat{y} = \underset{y \in \{1,2\}}{\arg\max} \left(\log P(\tilde{y} = y) + \sum_{k=1}^{K} \tilde{\rho}_k \log(p_{k,y}) \right). \tag{9}$$

As the MNB classifier is linear in the log-space, we can easily interpret the decision boundary. If a new observation $\tilde{\boldsymbol{x}}$ takes on a state which was not observed during the training phase, $p_{k,\tilde{y}}$ will be zero, resulting in setting the whole probability estimate to zero. Additive smoothing techniques [26], in which a small sample-correction is added to all probability estimates, are popular approaches to prevent the probability estimate to be zero. In [14], an unobserved state indicator is classified as protocol violation. In this paper, we take advantage of the MFCC's properties to tackle this problem. We assume that the MFCCs of the signals with similar characteristics have similar distributions [25]. Then, by calculating the Mahalanobis distance between the MFCCs of the new observation (whose state indicator value has not been seen during the training phase) and the center of all the observed clusters, we replace $p_{k,\tilde{y}}$ by the probability of the cluster which is closest to the observation.

3. EXPERIMENTAL SETUP

The proposed method has been validated on voice recordings from the Smartphone-PD study [6] collected by smartphones. This data set contains more than 7,500 recordings of 20 second sustained vowel /a/ phonations collected via an Android smartphone application by PD patients and healthy controls from all over the world. The designed voice test protocol for this database is described in Section 2.1. To evaluate the performance of the proposed approach, ground truth labels are needed. To this end, we first selected a subset of 100 recordings (50 PD patients and 50 healthy controls equally from both males and females) uniformly at random so as to have a reasonably large population that is practical to annotate the frames manually. The hand labeling was performed by playing back the recordings in the Audacity software and annotating the frames according to whether they adhered or violated the test protocol. Using a Hamming window, recordings are segmented into frames of 30 ms with 10 ms overlap. For each frame of a signal, 12 MFCCs along with the log energy are calculated. The features of every ten consecutive frames are averaged to smooth out the impact of articulation [25], and to prevent capturing very small changes in signal characteristics,

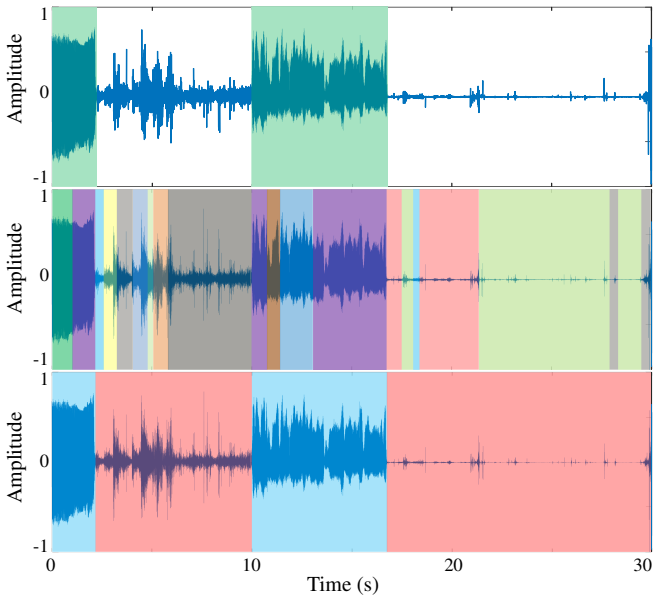

Fig. 2: Illustrative results of applying the proposed method to a 30-second segment of the voice recordings selected from the data set. The green shaded areas in the top plot represent the segments of the signal which are hand-labeled as adhering to the protocol. The middle plot shows the states, generated by the iHMM, in different colors. The bottom plot illustrates the result of applying a trained classifier to the state indicators to predict which segments adhere to (shaded in blue) and which ones violate the protocol (shaded in red).

which results in producing many uninterpretable states. Thus, each observation represents an averaged MFCCs of ≈ 200 ms of a signal. For the iHMM, we used the conjugate normal-gamma prior over the Gaussian state parameters, set the hyper-parameters $\alpha = 10$ and $\gamma = 10$, and run the inference for 150 iterations.

4. RESULTS

All the voice recordings in the data set are normalized with respect to their maximum absolute amplitude and concatenated to form one large recording of length 2,000 seconds. The top plot in Fig. 2 shows a segment of 30 seconds duration selected from the data set. The segments of the signal which adhere to the test protocol are hand-labeled and shaded in green. Fitting the iHMM to the data, the Gibbs sampler converged after tens of iterations (verified by inspecting the joint data log-likelihood), and 48 different states were discovered in this particular subset, each of them representing a different event in the signal. We observed that changes in signal characteristics (due to, e.g., voice disorders, change in pitch, talking, laughing, coughing, and different types of signal degradation) result in producing different states. This facilitates discovering different phenomena in the voice recordings. The middle plot in Fig. 2 illustrates the generated states in different colors. To evaluate the performance of the proposed approach in identifiying segments of the recordings that adhere or violate the test protocol in data not observed during the training phase (i.e. out of sample), we used 10-fold cross validation (CV) in which the recordings were randomly divided into 10 non-overlapping and equal sized subsets (10 recordings per subset). Since the MNB classifier requires a small amount of training data to estimate the parameters of the decision boundary, it is enough to use a single partition for training and validate on the remaining data.

Table 1: Comparison of the baseline systems and the proposed method for quality control applied to the recordings of the Smartphone-PD study. Results are in the form of mean $\pm$ STD computed using 10-fold CV and 200 repetitions. Since the VAD-based method is unsupervised, we used all data for evaluation; thus, the STD is not reported as it is not meaningful for a single trial.

Method	TPR	TNR	Accuracy
VAD-Based	84%	96%	90%
NPSAR-Based	88% $\pm$ 9%	91% $\pm$ 9%	89% $\pm$ 8%
Proposed	97% $\pm$ 2%	96% $\pm$ 4%	96% $\pm$ 2%

This process is repeated 10 times so that all subsets are used once for training the model. We repeated the CV procedure 200 times to obtain the distribution of classification accuracies.

We compare the proposed iHMM approach trained in the MFCC domain with two different baseline methods: the energy-based VAD which computes the energies of all frames, selects the maximum, and then sets the detection threshold as 30 dB below the maximum [27]; and the nonparametric switching AR model proposed in [14] which takes as an input the energy of each frame. We refer to the former baseline as the VAD-based method and to the latter one as NPSAR-based method in the rest of the paper. Table 1 shows the results of the baseline systems and the proposed method over all CV repetitions in terms of the true positive rate (TPR), true negative rate (TNR), and overall classification accuracy defined respectively as:

$$\text{TPR} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}, \tag{10}$$

$$\text{TNR} = \frac{\text{true negative}}{\text{true negative} + \text{false positive}}, \tag{11}$$

$$\text{Accuracy} = \frac{\text{true positive} + \text{true negative}}{\text{total number of test samples}}. \tag{12}$$

The results show that the proposed approach outperforms both baseline methods. It can be observed that many useful frames have not been detected by the baseline methods. Moreover, we expect a lower TNR value for the VAD-based approach when the data set contains more high energy protocol violations such as talking, laughing or high level of background noise. The bottom plot in Fig. 2 illustrates an example result of mapping the state indicators to adherence vs violation labels using a trained classifier. This remarkable performance, compared with the baseline methods, could mainly be due to fitting an iHMM to a richer feature set (MFCCs) in contrast to energy of the frames as used in both baselines. The results also suggest that with a small amount of hand-labeled data, the proposed method can automatically detect segments of the voice recordings that adhere to the test protocol from a large data set with a 0.2 s resolution and high accuracy. We also observed that a higher temporal resolution in identifying protocol violations can be achieved by increasing the amount of training data.

5. CONCLUSION

In this paper, we have proposed a new approach for implementing automatic quality control on voice recordings collected for remotely monitoring PD voice symptoms. This method is based on splitting signals into variable duration segments by fitting an infinite HMM to the frames of the signals in MFCC domain, and subsequently identifying segments that adhere to the voice test protocols by applying a simple and highly interpretable classifier. Using a small amount of hand-labeled data, the proposed approach can achieve a high accuracy (96%) in detecting short-term protocol violations with a 0.2 second resolution.

6. REFERENCES

[1] A. Zhan, M. A. Little, D. A. Harris, S. O. Abiola, E. R. Dorsey, S. Saria, and A. Terzis, "High Frequency Remote Monitoring of Parkinson's Disease via Smartphone: Platform Overview and Medication Response Detection," *arXiv:1601.00960v1*, 2016.

[2] M. Shahbakhi, D. T. Far, and E. Tahami, "Speech analysis for diagnosis of Parkinson's disease using genetic algorithm and support vector machine," *Journal of Biomedical Science and Engineering*, vol. 07, no. 04, pp. 147–156, 2014.

[3] D. Gil and M. Johnson, "Diagnosing Parkinson by using artificial neural networks and support vector machines," *Global Journal of Computer Science and Technology*, pp. 63–71, 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 Transactions on Biomedical Engineering*, vol. 59, pp. 1264–1271, 2012.

[5] B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E. R. Dorsey, S. H. Friend, and A. D. Trister, "The mPower study, Parkinson disease mobile data collected using ResearchKit," *Scientific Data*, vol. 3, no. 160011, 2016.

[6] A. Zhan, M. A. Little, D. A. Harris, S. O. Abiola, E. R. Dorsey, S. Saria, and A. Terzis, "High frequency remote monitoring of Parkinson's disease via smartphone: platform overview and medication response detection," pp. 1–12, 2016.

[7] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *National Science Review*, vol. 1, no. 2, pp. 293–314, jun 2014.

[8] H. E. Plesser, "Reproducibility vs. replicability: A brief history of a confused terminology," *Frontiers in neuroinformatics*, vol. 11, p. 76, 2017.

[9] M. Fakhry, A. H. Poorjam, and M. G. Christensen, "Speech enhancement by classification of noisy signals decomposed using NMF and Wiener filtering," in *26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, 2018.

[10] J. Ramirez, J. M. Gorriz, and J. C. Segura, "Voice activity detection. Fundamentals and speech recognition system robustness," in *Robust Speech Recognition and Understanding*, 2007, pp. 1–22.

[11] C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno, T. Power, A. Sahuguet, M. Shugrina, and O. Siohan, "An audio indexing system for election video material," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 4873–4876.

[12] A. H. Poorjam, J. R. Jensen, M. A. Little, and M. G. Christensen, "Dominant distortion classification for pre-processing of vowels in remote biomedical voice analysis," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 289–293.

[13] A. H. Poorjam, M. A. Little, J. R. Jensen, and M. G. Christensen, "A parametric approach for classification of distortions in pathological voices," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018.

[14] R. Badawy, Y. Raykov, L. Evers, B. Bloem, M. Faber, A. Zhan, K. Claes, M. Little, R. Badawy, Y. P. Raykov, L. J. W. Evers, B. R. Bloem, M. J. Faber, A. Zhan, K. Claes, and M. A. Little, "Automated quality control for sensor based symptom measurement performed outside the lab," *Sensors*, vol. 18, no. 4, 2018.

[15] K. Knill and S. Young, "Hidden Markov models in speech and language processing," in *Corpus-Based Methods in Language and Speech Processing*. Springer, 1997, pp. 27–68.

[16] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[17] S. L. Scott, "Bayesian methods for hidden Markov models: Recursive Computing in the 21st Century," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 337–351, 2002.

[18] A. Stolcke and S. Omohundro, "Hidden Markov model induction by Bayesian model merging," in *Advances in Neural Information Processing Systems*, vol. 5, 1993, pp. 11—-18.

[19] M. J. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, University College London, 2003.

[20] C. E. Rasmussen, "The infinite Gaussian mixture model," *Advances in Neural Information Processing Systems 12*, pp. 554–560, 2000.

[21] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen, "The infinite hidden Markov model," *Advances in Neural Information Processing Systems 14*, pp. 577–584, 2002.

[22] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[23] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[24] B. J. Mohan and R. Babu. N, "Speech recognition using MFCC and DTW," in *International Conference on Advances in Electrical Engineering (ICAEE)*, 2014, pp. 1–4.

[25] A. H. Poorjam, M. A. Little, J. R. Jensen, and M. G. Christensen, "A supervised approach to global signal-to-noise ratio estimation for whispered and pathological voices," in *ICASSP*, Calgary, Canada, 2018.

[26] Q. Yuan, G. Cong, and N. M. Thalmann, "Enhancing naive Bayes with various smoothing methods for short text classification," in *International conference on World Wide Web*, New York, USA, 2012, p. 645.

[27] K.-H. Woo, T.-Y. Yang, K.-J. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.