Quality Control in Remote Speech Data Collection

Amir Hossein Poorjam, *Student Member, IEEE*, Max A. Little, *Member, IEEE*, Jesper Rindom Jensen, *Member, IEEE*, and Mads Græsbøll Christensen, *Senior Member, IEEE*

***Abstract*—There is the need for algorithms that can automatically control the quality of the remotely collected speech databases by detecting potential outliers which deserve further investigation. In this paper, a simple and effective approach for identification of outliers in a speech database is proposed. Using the deterministic minimum covariance determinant (DetMCD) algorithm to estimate the mean and covariance of the speech data in the mel-frequency cepstral domain, this approach identifies potential outliers based on the statistical distance of the observations in the feature space from the central location of the data that are larger than a predefined threshold. The DetMCD is a computationally efficient algorithm which provides a highly robust estimate of the mean and covariance in multivariate data even when 50% of the data are outliers. Experimental results using 8 different speech databases with manually inserted outliers show the effectiveness of the proposed method for outlier detection in speech databases. Moreover, applying the proposed method to a remotely collected Parkinson's voice database shows that the outliers that are part of the database are detected with 97.4% accuracy, resulting in significantly decreasing the effort required for manually controlling the quality of the database.**

***Index Terms*—Outlier detection, quality control, robust estimation, speech database, remote data collection.**

I. Introduction

Development of many speech-based systems such as speech and language recognition, speaker characterization and biomedical speech analysis requires a large amount of high-quality training data to accurately model the characteristics of the speech signals [1]–[6]. In many cases, such databases do not exist, and therefore, speech samples should be collected prior to development of a system. The acquisition of speech data, in a broad sense, can be separated into two categories, namely supervised and unsupervised data collection. In supervised speech data collection, participants are typically required to be present on-site to record their voice in specific experimental conditions. The process of data collection is supervised and controlled by an expert to train participants, verify the equipment is correctly configured, check that recordings are complete, and repeat the procedure if a recording does not satisfy the requirements [7]. While high quality samples can be collected under controlled conditions, creating large speech databases this way is challenging and impractical in most cases.

With unsupervised speech data collection on the other hand, a speech database is created remotely to bypass the logistical limitations of controlled approaches. In this method, participants are typically provided with a web-based interface, a dedicated line to make a phone call or an application running on remote devices such as smartphones, tablets or personal computers accompanied by instructions on how to record and submit voice recordings. Participants can contribute to speech data collection at any time and in any location. SWITCHBOARD is an example of a large multispeaker database of more than 250 hours of conversational speech collected automatically over telephone lines by 500 speakers from around the US [8]. The LibriVox project [9] is another example large speech database which contains more than 1,000 hours of recordings collected by volunteers reading chapters of books and submitting the recordings to a web server using a web-based interface. Lane et al. developed two tools to collect speech data remotely using mobile devices and via a web-based interface [10]. Using an online educational game called Voice Scatter, Gruenstein et al. [11] collected more than 27 hours of speech data remotely from 1193 speakers in 22 days through a web-based interface. Nagrani et al. [12] used a multi-stage approach to automatically collect a large scale speaker recognition database, called VoxCeleb, from YouTube videos of 1,251 celebrities. Although remote data collection is easier and results in a larger population sample compared to supervised collection, the quality of the recordings is often poor and the homogeneity of the recordings in the database is not guaranteed, because speech samples can be recorded in a wide range of environments using different recording devices. In this case, the amount of usable recordings depends upon how well participants are trained before starting the recording procedure [10]. Even though the quality of recordings can be controlled prior to submission by playing back the recorded samples and repeating the recording procedure in case the speech signals do not satisfy requirements [13], some participants still submit defective recordings due to lack of training, misinterpretation of protocols or negligence. Moreover, not all interfaces facilitate playing back the audio signals prior to submission [14], [15]. The presence of inconsistent and/or low-quality samples in a speech database can significantly degrade the performance of speech-based applications. Therefore, speech databases typically need to be analyzed and cleaned before being processed. Controlling the quality of recordings is one of the major challenges in speech database collection and it is typically performed by experts. As human inspection is often infeasible for large databases which contain hundreds of hours of recordings, there is the need for automatic methods to recognize low-quality samples in speech databases. The European Language Resources Association

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

A. H. Poorjam, J. R. Jensen and M. G. Christensen are with the Audio Analysis Lab, CREATE, Aalborg University, Aalborg 9000, Denmark (e-mail: ahp@create.aau.dk; jrj@create.aau.dk; mgc@create.aau.dk).

M. A. Little is with the School of Engineering and Applied Science, Aston University, Birmingham, UK and also with Media Lab, MIT, Cambridge, Massachusetts, USA (e-mail: max.little@aston.ac.uk).

(ELRA) considers the signal-to-noise ratio (SNR), clipping rate and mean amplitude of the recordings as measures for a quick quality control of the audio signals in a speech database [16]. However, the performance of these measures is limited when searching for recordings that are of high-quality but varying in terms of content and recordings with very short duration speech activities. Moreover, we have shown in [17] that most existing SNR estimation methods only work for normal speech signals and cannot provide accurate estimation for special speech types such as whispered or disordered voices.

The process of post-hoc quality control in speech databases can be considered as an outlier detection problem in which the samples that, in some sense, are "far" from the majority of the data that have been collected for a particular purpose are considered as outliers. However, the definition of outliers in speech databases differs across applications. In this paper, we aim to identify outliers which are produced during the remote data collection process such as errors due to technical problems in recording equipment, participants' mistakes during recording or recording in very noisy acoustic environments. We propose a simple and efficient approach to finding low-quality and inconsistent samples in a speech database by identifying statistical outliers. In this method, a deterministic minimum covariance determinant algorithm (DetMCD) [18] is used to provide robust estimation of mean and covariance of the mel-frequency cepstral coefficients (MFCCs) extracted from recordings. Outliers, which may deserve additional inspection, are then determined by selecting samples whose robust statistical distance in feature space from the robust central location of the data is larger than a predefined threshold.

The paper is organized as follows. In Section II, the outlier detection problem in a speech database is formulated and the algorithms in our proposed method are explained. The proposed method is elaborated in Section III. Section IV introduces the experimental setup and describes the databases used in this study. In Section V, the experimental results are presented. Finally, Section VI summarizes the paper.

II. Background

A. Problem Formulation

Controlling the quality of recordings in a speech database can be considered as the identification of the potential outliers in a database. In general, depending on the amount of additional information provided for the recordings, a speech database can be used for many different applications such as speech recognition, emotion classification, voice disorder diagnosis, speaker recognition, language recognition and environmental sniffing. Consequently, outliers in a specific speech database can differ across applications. However, in this paper, our main concern is to identify outliers occurring during the remote data collection process.

Errors in large speech databases can occur at every step of the remote data collection process. An understanding of the types of errors occurring during data collection facilitates the development of an appropriate automatic data clean-up method. Although not exhaustive, the major types of errors fall into one or more of the following categories:

- **Empty or very short speech activity:** This type of error, in which there is no or a very short speech activity in a signal (comparing to the signal length), can happen due to technical problems in recording equipment or participants' mistakes during recording. In this case, the speech sample is useless since no relevant information can be captured from the sample.
- **Low-quality samples:** This type of error can occur due to poor or misconfigured recording equipment, recording in a noisy or reverberant acoustic environment, and common processing through nonlinear elements such as an audio codec or hard clipping.
- **Wrong context:** Recordings in this case can even be of high-quality but they do not comply with the context of the database. For example, the presence of a non-speech sound or a recording of whispered speech are considered as incorrect data in a normal voice speech database. This error can typically originate from participants' mistakes due to misinterpretation of the speech/voice task, submitting an incorrect speech sample or saving a submitted recording in the wrong repository on the web-server.

For outlier detection in a speech database, we are given a set of data, $\mathcal{Z} = \{\boldsymbol{z}_i\}_{i=1}^{n}$, where $\boldsymbol{z}_i$ denotes the i^{th} recording in the database. The goal is to identify as many low-quality and inconsistent samples as possible which are outliers with respect to the majority of samples in the database. Depending on the application, a flagged outlier can either be kept in the database if it has been a false alarm, be enhanced and kept in the database if it is degraded [19], or be excluded from the database if it is not possible to retrieve useful information from the signal.

B. Robust Mean and Covariance Estimators

Given an $n \times m$ matrix $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n]^T$ with the i^{th} observation $\boldsymbol{z}_i = [z_{i1}, \ldots, z_{im}]^T$ of dimension m, the center and scatter of the data set are typically estimated by calculating the sample mean and the sample covariance. However, these estimates are highly affected by the presence of outliers in a data set. They have a breakdown value—the smallest percentage of outliers which can have an arbitrarily large impact on the estimator—of $1/n$ which means that even a single outlier can modify both estimates arbitrarily.

To make the multivariate mean and covariance estimators robust against outliers, Rousseeuw proposed to find a subset of k observations, where $\frac{n}{2} \leq k \leq n$, which has the minimum volume ellipsoid (MVE) [20]. This subset is then used to estimate the robust mean and covariance of the data set. Although the MVE has a high breakdown value, $\approx 50\%$, its slow convergence rate makes this algorithm inefficient. The minimum covariance determinant (MCD) method [21] is an alternative which provides a highly robust estimate of the mean and covariance in multivariate data. This approach looks for a subset of k samples for which the covariance matrix has the lowest determinant. The mean and the covariance of this subset is then used as the estimate of mean and covariance of the data set. Even though it provides higher accuracy than the MVE and has a breakdown value of 50% when $k = \lfloor (m+n+1)/2 \rfloor$

[22], the MCD is computationally inefficient. In the Fast-MCD algorithm [22], Rousseeuw and van Driessen proposed an approximation to the MCD by incorporating a fast re-sampling algorithm to select a large number of random subsets and applying concentration steps to select subsets whose covariances have the smallest determinant. Since Fast-MCD starts by drawing initial random subsets, the results are not necessarily the same at multiple runs of the algorithm. As an alternative approximation to MCD, Hubert et al. proposed a deterministic algorithm (DetMCD) [18] to compute the MCD estimator. This algorithm starts from only a few well chosen initial estimates which are used to form six initial robust subsets. Then, the concentration steps are applied to these subsets until convergence. Since our proposed method is based on the DetMCD algorithm, we describe this algorithm in more detail below.

In DetMCD, each column of the data matrix $\boldsymbol{Z}_j$ (j=$1,\ldots,m$) is first subtracted by its median and divided by the scale estimator $Q_n(\boldsymbol{Z}_j)$=$2.2219\times\{|z_{cj}-z_{dj}|; c<d\}_{(p)}$ which is the p^{th} order statistic of the $\binom{n}{2}$ interpoint distances [23], where p=$\binom{h}{2}$ and h=$[n/2]$+1. This standardization makes the algorithm location and scale equivariant. Given the standardized data matrix $\boldsymbol{X}$ with rows $\boldsymbol{x}_i^T$ ($i=1,\ldots,n$) and columns $\boldsymbol{X}_j$ ($j=1,\ldots,m$), six initial estimates $\boldsymbol{S}_l$ ($l=1,\ldots,6$) of the correlation or covariance matrix of $\boldsymbol{X}$ are constructed [18] as follows:

- $\boldsymbol{S}_1=\operatorname{corr}(\boldsymbol{W})$ with $\boldsymbol{W}_j=\tanh(\boldsymbol{X}_j)$, where $\boldsymbol{W}_j$ ($j=1,\ldots,m$) are the columns of $\boldsymbol{W}$.
- $\boldsymbol{S}_2=\operatorname{corr}(\boldsymbol{R})$ where $\boldsymbol{R}$, with columns $\boldsymbol{R}_j$ (j=$1,\ldots,m$), is the rank of $\boldsymbol{X}$. The matrix $\boldsymbol{S}_2$ is the Spearman correlation matrix of $\boldsymbol{X}$.
- $\boldsymbol{S}_3=\operatorname{corr}(\boldsymbol{T})$ with $\boldsymbol{T}_j=\Phi^{-1}\big((\boldsymbol{R}_j-\frac{1}{3})/(n+\frac{1}{3})\big)$, where $\Phi(\cdot)$ is the normal cumulative distribution function, and $\boldsymbol{T}_j$ ($j=1,\ldots,m$) are the columns of $\boldsymbol{T}$.
- $\boldsymbol{S}_4=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{k}_i\boldsymbol{k}_i^T$ with $\boldsymbol{k}_i\overset{\text{def}}{=}\boldsymbol{x}_i/\|\boldsymbol{x}_i\|$ for all i.
- $\boldsymbol{S}_5=\operatorname{cov}(\boldsymbol{Y})$, where $\boldsymbol{Y}$ is the $\lceil n/2\rceil$ standardized observations $\boldsymbol{x}_i$ with smallest norm.
- $\boldsymbol{S}_6$ is the raw orthogonalized Gnanadesikan-Kettenring (OKG) estimator [24].

Then, for each of these estimates $\boldsymbol{S}_l$, the Mahalanobis distance of the observations are calculated as:

$$\mathrm{MD}_{il}=D\left(\boldsymbol{x}_i,\hat{\boldsymbol{\mu}}_l,\hat{\boldsymbol{\Sigma}}_l\right)=\sqrt{\left(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_l\right)^T\hat{\boldsymbol{\Sigma}}_l^{-1}\left(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_l\right)},\tag{1}$$

where the covariance and the center of $\boldsymbol{X}$ are estimated using

$$\hat{\boldsymbol{\Sigma}}_l=\boldsymbol{E}\boldsymbol{L}\boldsymbol{E}^T,\tag{2}$$

$$\hat{\boldsymbol{\mu}}_l=\hat{\boldsymbol{\Sigma}}_l^{-1/2}\left(\operatorname{comed}\left(\boldsymbol{X}\hat{\boldsymbol{\Sigma}}_l^{-1/2}\right)\right),\tag{3}$$

in which $\boldsymbol{E}$ is the orthogonal matrix of eigenvectors of $\boldsymbol{S}_l$, $\boldsymbol{L}$=$\operatorname{diag}\big(Q_n^2(\boldsymbol{V}_1),\ldots,Q_n^2(\boldsymbol{V}_m)\big)$ with $\boldsymbol{V}$=$\boldsymbol{X}\boldsymbol{E}$, and $\operatorname{comed}(\cdot)$ denotes the coordinate-wise median.

In the next step, the mean and covariance matrix of the $k_0=\lceil n/2\rceil$ observations with smallest MD_{il} are computed for each initial estimate l, and the new statistical distances (denoted as MD_{il}^*) for all n observations are calculated. Then, k observations with smallest MD_{il}^* are selected for each $l=1,\ldots,6$ and the concentration step is applied to them until convergence.

In the concentration step, the statistical distances $d_{\text{old}}(i)$=$D(\boldsymbol{z}_i,\hat{\boldsymbol{\mu}}_{\text{old}},\hat{\boldsymbol{\Sigma}}_{\text{old}})$ for all n observations are computed given the initial estimates of the mean $\hat{\boldsymbol{\mu}}_{\text{old}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_{\text{old}}$. By sorting these distances, a permutation τ for which $d_{\text{old}}(\tau_1)\le d_{\text{old}}(\tau_2)\le\cdots\le d_{\text{old}}(\tau_n)$ is obtained. The new estimates of the mean $\hat{\boldsymbol{\mu}}_{\text{new}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_{\text{new}}$ are respectively computed as

$$\hat{\boldsymbol{\mu}}_{\text{new}}=\frac{1}{k}\sum_{i\in K}\boldsymbol{z}_i,\tag{4}$$

$$\hat{\boldsymbol{\Sigma}}_{\text{new}}=\frac{1}{k-1}\sum_{i\in K}\big(\boldsymbol{z}_i-\hat{\boldsymbol{\mu}}_{\text{new}}\big)\big(\boldsymbol{z}_i-\hat{\boldsymbol{\mu}}_{\text{new}}\big)^T,\tag{5}$$

where $K=\{\tau_1,\tau_2,\ldots,\tau_k\}$. It was proved in [22] that the determinant of $\hat{\boldsymbol{\Sigma}}_{\text{new}}$ is smaller than or equal to the determinant of $\hat{\boldsymbol{\Sigma}}_{\text{old}}$ with equality only if $\hat{\boldsymbol{\Sigma}}_{\text{new}}=\hat{\boldsymbol{\Sigma}}_{\text{old}}$, which means that the sequence of determinants converges in a finite number of steps.

Finally, a weighting step is applied to increase the statistical efficiency of the estimated mean and covariance matrix as:

$$\tilde{\boldsymbol{\mu}}=\frac{\sum_{i=1}^{n}\rho(d_i^2)\boldsymbol{z}_i}{\sum_{i=1}^{n}\rho(d_i^2)},\tag{6}$$

$$\tilde{\boldsymbol{\Sigma}}=\frac{1}{\sum_{i=1}^{n}\rho(d_i^2)-1}\sum_{i=1}^{n}\rho(d_i^2)(\boldsymbol{z}_i-\tilde{\boldsymbol{\mu}})(\boldsymbol{z}_i-\tilde{\boldsymbol{\mu}})^T,\tag{7}$$

with weights

$$\rho(d_i^2)=\begin{cases}1 & d_i^2\le\chi^2_{m,\alpha}\\0 & \text{otherwise},\end{cases}\tag{8}$$

where $\chi^2_{m,\alpha}$ is the α-quantile of the Chi-square distribution with m degrees of freedom.

The permutation invariant property of DetMCD makes the results independent of the order of the observations in the data set. Using DetMCD, it is recommended to have subsets of $k\approx0.5n$ when the data set is expected to contain many outliers and $k\approx0.75n$ otherwise [18].

C. Robust Statistical Distance

Outliers in a data set can be considered as the observations that, in some sense, are "far" from the rest of the data. The Mahalanobis distance, which is defined in (1), is a widely used metric for measuring the distance between an observation and the center of a distribution that takes the covariance of the distribution into account. The Mahalanobis distance is a useful metric for detecting a single outlier in a data set [25]. However, it is not robust against outliers, particularly when multiple outliers are present in a data set, since the sample mean and the sample covariance estimates in (1) are sensitive to outliers. Replacing these estimates in the Mahalanobis distance by the robust estimates of mean $\tilde{\boldsymbol{\mu}}$ and covariance $\tilde{\boldsymbol{\Sigma}}$, computed in (6) and (7) respectively, the robust distance is defined as:

$$\mathrm{RD}_i=D\left(\boldsymbol{z}_i,\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}}\right)=\sqrt{\left(\boldsymbol{z}_i-\tilde{\boldsymbol{\mu}}\right)^T\tilde{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{z}_i-\tilde{\boldsymbol{\mu}}\right)}.\tag{9}$$

III. The Proposed Method

The block diagram of the proposed method for detecting potential outliers in speech databases is illustrated in Fig.1. To identify outliers, we need features to reflect changes in signal characteristics at the recording level. In this paper, we propose to use mel-frequency cepstral coefficients (MFCCs) since not only do they convey information about speech context [26], but it has also been demonstrated that the presence of noise and distortion in speech signals predictably modify the distribution of the MFCCs by changing the covariance of MFCCs and shifting the mean to different regions in the feature space [17], [27]. Moreover, we have shown in [17] that if the MFCCs of a recording are averaged over frames, the amount of change in the mean and covariance matrix of the MFCCs is related to the level of overall noise and distortion in the recording regardless of the speech type. Thus, assuming that the majority of recordings in a speech database are of, more or less, the same quality and there is no context variability among them, we expect the corresponding MFCCs to have a smaller distance from the central location of the data compared to the MFCCs of the potential outliers, which we assume to have a different distribution to the rest of the data. To this end, the MFCCs extracted from the frames of a recording are averaged to form a fixed-length, low-dimensional vector per recording. Using the DetMCD algorithm to estimate the robust mean and covariance of the data, the robust statistical distance, RD, for all observations in the data set is calculated using (9). Then, the observations with a robust distance larger than a predefined threshold are expected to be the potential outliers that deserve additional inspection.

Setting a threshold is to some extent arbitrary, database-dependent and requires domain knowledge. However, since the asymptotic distribution of the robust distances is the Chi-square (χ^2_m) distribution with m degrees of freedom [28], the RD values larger than the threshold, defined as

$$\theta = \sqrt{\chi^2_{m,0.975}}\ , \tag{10}$$

can be considered as potential outliers. Here $\chi^2_{m,\alpha}$ is the α-quantile of the Chi-square distribution. The cut-off value depends on α and the number of variables m. Choosing a small value for α leads to flagging too many observations as outlying (false positives or type I errors). On the other hand, setting α to a value very close to one results in missing the potential outliers (false negatives or type II errors). We propose to set $\alpha = 0.975$. The impact of the number of variables on the performance of the algorithm is investigated in Section V.

IV. Experimental Setup

The proposed approach has been validated on eight databases of three different speech types, namely normal speech, whispered speech and pathological voice. Specifically, for normal speech, the LibriSpeech database [9], the TIMIT continuous speech database [7], a noisy speech database [29], and a noisy reverberant speech database [30] have been used. The LibriSpeech database is based on the LibriVox project[1] containing more than 1,000 hours of audio books read in English by both male and female speakers. The TIMIT continuous speech database has been developed for evaluation of automatic speech recognition systems and contains 6,300 utterances uttered by 630 English speakers of both genders. The noisy speech and noisy reverberant speech databases were created based on the clean recordings of the Voice Bank corpus [31] and designed to train and test speech enhancement methods and text-to-speech models. To create the noisy database, clean recordings of 14 male speakers and 14 female speakers were contaminated by ten different types of noise at 10 dB and 15 dB. For the noisy reverberant database, the clean recordings of 14 male speakers and 14 female speakers were made reverberant by convolving them with room impulse responses of three different databases, and adding them to the noisy signals at 10 dB and 15 dB. The last two databases have been selected because there is a growing demand for collecting noisy and revereberant speech databases that are similar to live recordings with a microphone.

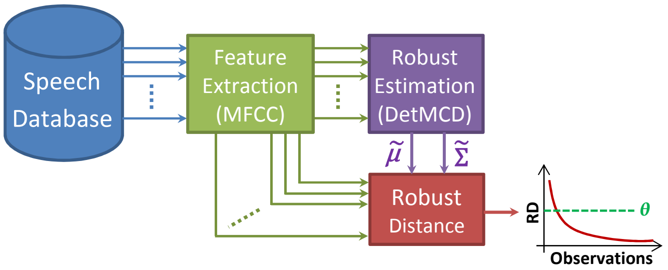

Fig. 1. Block diagram of the proposed method for identification of potential low-quality and inconsistent outliers in a speech database.

Whispering is often used for quiet or private communication. To evaluate the performance of the proposed algorithm on the whispered speech databases, we considered the CHAIN database [32] and the CSTR NAM TIMIT Plus database [33]. The CHAIN database contains whispered speech samples uttered by 36 English speakers of both genders. The CSTR NAM TIMIT Plus database consists of 420 sentences read with a whispered voice and recorded using an omni-directional headset-mounted condenser microphone. The recordings of both databases have been collected in noise free environments.

Due to the development of advanced machine learning techniques, many voice disorders can be deteced using voice signals [5], [34], [35]. To develop accurate and reliable algorithms for detection of disorders from voice signals a very large number of good- and consistent-quality voice recordings are required. In this study, we considered two voice databases from Parkinson's disease patients as examples of remotely collected pathological voice databases. The first database, generated through collaboration between Sage Bionetworks, PatientsLikeMe and Dr. Max Little as part of the Patient Voice Analysis study[2], includes telephone recordings of the sustained vowel /a/ uttered by 750 patients of both genders. The second database consists of more than 65,000 samples of the sustained vowel /a/ recorded via smartphones by healthy and patient speakers of both genders. This database has been developed through the mPower mobile Parkinson Disease study [15] in

[1] https://librivox.org

[2] Obtained through Synapse ID [syn2321745]

TABLE I
SUMMARY DETAILS OF THE DATABASES

Database	Speech Type	Duration (Min./Avg./max.)	Sampling Rate	Gender	Collection Method
LibriSpeech	Normal Speech	9.6 / 15 / 34 sec.	16 kHz	Male/Female	Remotely by Mic.
TIMIT Continuous Speech	Normal Speech	5 / 5.5 / 7.8 sec.	16 kHz	Male/Female	On-site by Mic.
Noisy Speech	Normal Speech	5 / 6.5 / 12 sec.	48 kHz	Male/Female	On-site by Mic.
Noisy Reverberant Speech	Normal Speech	2.3 / 8 / 15 sec.	48 kHz	Male/Female	On-site by Mic.
CHAIN	Whispered Speech	1.5 / 7.5 / 57.5 sec.	44.1 kHz	Male/Female	On-site by Mic.
CSTR NAM TIMIT Plus	Whispered Speech	3.5 / 5.8 / 16.5 sec.	96 kHz	Female	On-site by Mic.
Telephone Parkinson's Database	Pathological Voice	3.1 / 16.5 / 29.5 sec.	8 kHz	Male/Female	Remotely by Telephone
Smartphone Parkinson's Database	Pathological Voice	9.9 / 10 / 10.1 sec.	44.1 kHz	Male/Female	Remotely by Smartphone

which participants from the US submit their voice using a mobile application when they have access to the internet. To distinguish between these two databases, we refer to the former as the telephone Parkinson's database and to the latter as the smartphone Parkinson's database in the rest of paper. Summary details of the databases are presented in Table I.

In order to evaluate the performance of the proposed method, ground truth labelling is needed. To this end, 200 consistent-quality recordings have been selected from each database and the following recordings, as examples of the most common outliers in speech databases, have been added to each database: (1) a silent signal, recorded in a very quiet room, with the same duration as the average duration of the recordings in the database; (2) a recording selected from the target database (the database under study), and the whole signal, save for a very short segment (100 ms), is set to silent: this outlier is considered as a recording with very short speech activity; (3) four recordings selected from the target database, two of them are contaminated by speech babble noise at 5 dB and -5 dB and the other two recordings are moderately and heavily distorted by clipping followed by reverberation: these four outlier recordings are used to represent low-quality recordings in a database; (4) four clean recordings from a different context selected from databases other than the one under analysis and one music signal played by a piano: these recordings are resampled and trimmed/repeated to have the same sampling rate and signal duration as those of the target database – these are considered as irrelevant samples (incorrect context) in the database. It should be noted that the four recordings from different databases have not been selected from the noisy speech nor from the noisy reverberant speech databases since we have already considered noisy and reverberant samples in the previous item; (5) one recording in the same context as the database under study but selected from a different database to represent a sample collected under a different acoustic environment or using different recording equipment. Thus, each database contains 212 recordings among which 12 recordings are known outliers.

V. Results and Discussion

The proposed method operates on mel-cepstral features. Recordings in each database are segmented into frames of 30 ms, with 10 ms overlap, using a Hamming window. For each frame of a speech signal, m cepstral coefficients are calculated. The MFCCs extracted from the frames of a recording are then averaged both to smooth out the impact of articulation and to form one m-dimensional vector per each recording. Since the flagged outliers are the only samples that are subject to further inspection (as explained in Section II-A), minimizing the misdetection error is more important than minimizing the false alarms. In Fig. 2, the number of misdetected outliers in the eight different databases are plotted as a function of m, the number of cepstral coefficients. The plots suggest that the best performance is obtained when m is set to a number between 5 and 14 which, in most cases, results in detecting most of the added outliers to the databases.

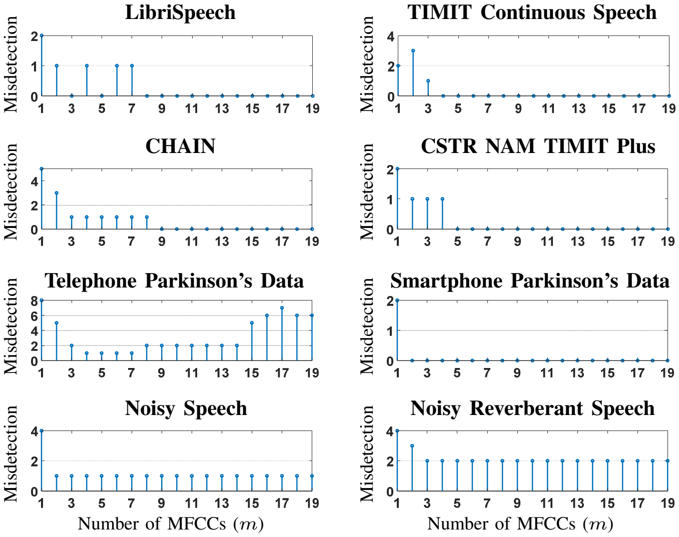

Fig. 2. Performance of the proposed method for outlier detection in speech databases, in terms of the number of misdetected outliers, as a function of the MFCC dimension.

The robust distance, RD, calculated from the observations of the eight speech databases using 5 MFCCs is shown in Fig. 3. In this figure, the threshold for identification of outliers are indicated by the dashed line. The correctly detected inlier observations which were collected according to the database's protocol are shown by the blue circles. The correctly detected outliers are represented by the green circles. The black stars show the misdetected outliers (false negatives), and the inlier samples detected as the potential outliers (false positives) are represented by the red crosses. Since most of the data in these plots are concentrated below the threshold, we show the vertical axes on a logarithmic scale for a better visualization. We set $k \approx 0.75n$ and $\theta = \sqrt{\chi^2_{5,0.975}} = 3.58$ for all databases. The results show the effectiveness of the proposed approach. We can observe that almost all outliers added to the databases

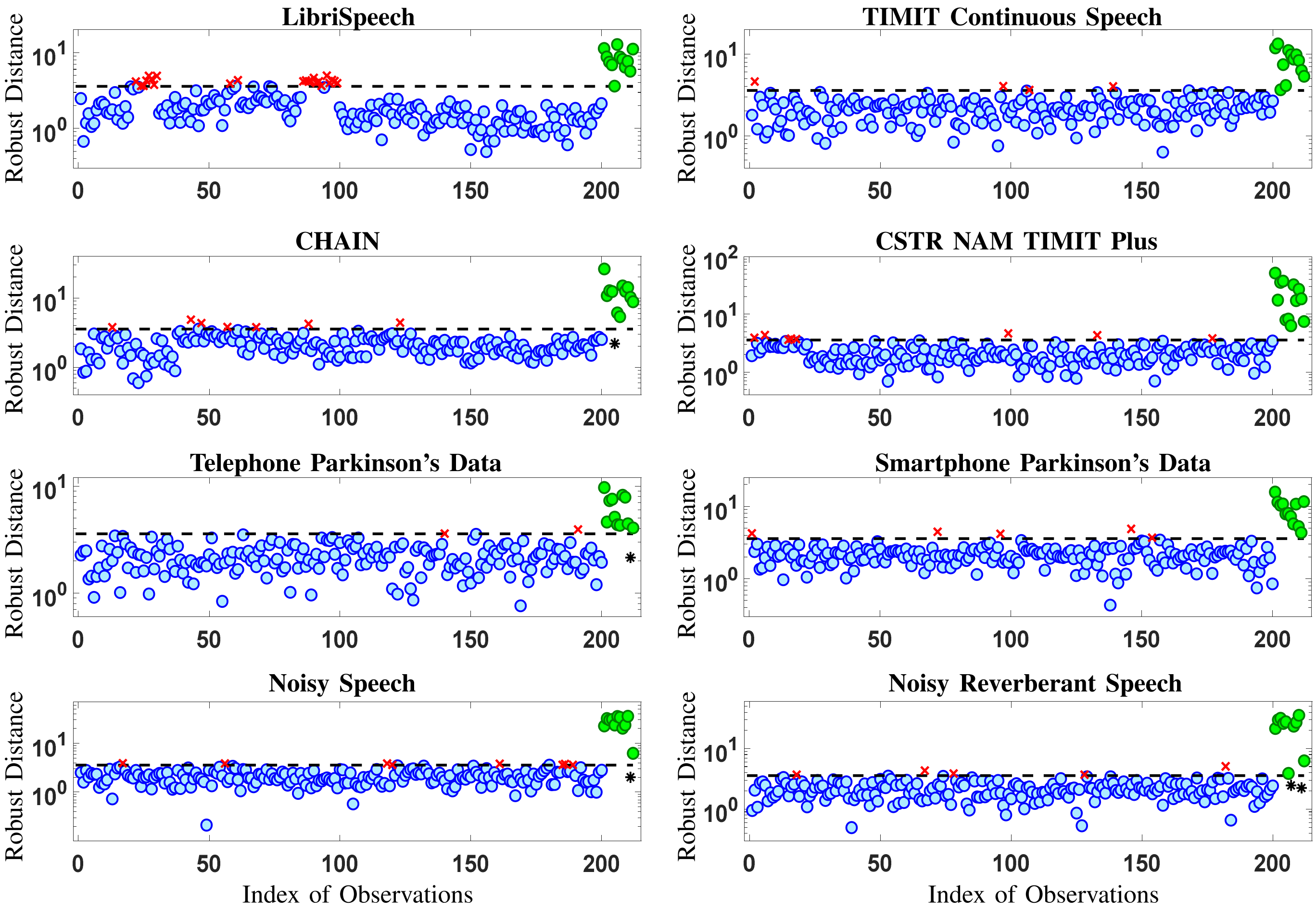

Fig. 3. Robust distance, on a logarithmic scale, calculated from the observations of eight different databases. The blue circles are the correctly detected inlier samples which are collected according to the corresponding data collection protocol. The green circles are the correctly detected outliers which are added to the databases. The red crosses indicate the inlier samples detected as the potential outliers (false positives). The black stars are the misdetected outliers (false negatives). The dashed lines indicate the threshold value for identification of outliers in databases defined in (10).

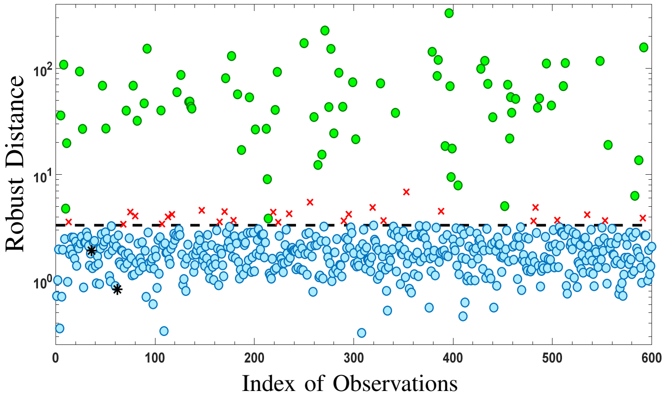

Fig. 4. Robust distance, on a logarithmic scale, calculated from the observations of a subset of 600 recordings randomly selected from the smartphone Parkinson's database. The blue circles are the clean observations correctly detected as inliers. The green circles are correctly detected outliers. The red crosses are the clean samples detected as the potential outliers (false positives). The black stars represent the misdetected outliers (false negatives). The dashed line shows the threshold value for identification of outliers.

have RD larger than the threshold defined in (10).

The outlier that is not detected in the CHAIN database is a recording in a whisper (the same context) added from the CSTR NAM TIMIT Plus database. The outlier which is not detected in the telephone Parkinson's database and the noisy speech database is a noisy recording at 5 dB, and two misdetected outliers in the noisy reverberant database are a noisy recording at 5 dB and a recording moderately distorted by clipping followed by reverberation. Bearing in mind that the recordings in the telephone Parkinson's database have been collected over the telephone, they are not perfectly clean and already have some noise and distortion. Moreover, the samples of the noisy speech and the noisy reverberant speech databases are the recordings that have been contaminated respectively by noise and noise-reverberation at 10 dB and 15 dB. Thus, recordings with moderate noise and distortion can still be considered as inliers in these particular databases. This suggests that noise and distortion in recordings can to some extent be tolerated depending on the data collection method.

In the next experiment, the proposed outlier detection method is applied to the smartphone Parkinson's database to evaluate the effectiveness of the proposed approach in detecting potential outliers in a real database which is remotely collected in an unsupervised manner. In this database, the

TABLE II
THE CONFUSION MATRIX OF THE PROPOSED METHOD APPLIED TO A SUBSET OF 600 RECORDINGS RANDOMLY SELECTED FROM THE SMARTPHONE PARKINSON'S DATABASE

		Predicted	
		Outlier	Inlier
Actual	Outlier	97.4 %	2.6 %
	Inlier	5.1 %	94.9 %

participants were supposed to utter a sustained vowel /a/ for 10 seconds at a comfortable pitch and intensity. Since the data collection was unsupervised, we expect to have outliers in the database. A subset of 600 recordings drawn uniformly at random from the database has been selected so as to have a reasonably large population while making it practical to annotate the recordings manually to assess the performance of the method. By manual inspection of the recordings, 77 outliers have been detected which can be categorized into three general classes, namely recordings with very short speech activity, recordings with no relevant speech activity (including speaking, laughing, coughing, recordings that captured only the ambient noise or empty recordings), and low quality recordings (noisy or distorted signals).

Fig. 4 shows the robust distance of the observations in this subset. The vertical axis is on a logarithmic scale for a better visualization of the data. The dashed line in this figure indicates the threshold. The blue circles show the good-quality samples which are correctly detected as inliers. Outliers correctly detected by the proposed algorithm are highlighted by the green circles. The good-quality samples detected as the potential outliers (the type I errors) are shown by the red crosses, and the misdetected outliers (the type II errors) are represented by the black stars. Setting $k \approx 0.75n$ and $\theta = 3.58$, the algorithm flagged 102 samples in this subset as potential outliers which have $\mathrm{RD} \geq \theta$ among which 75 samples are the actual outliers and 27 samples are good-quality recordings detected as outliers. Table II summarizes the results in the form of a confusion matrix. The results show that the proposed algorithm, for this particular subset, reduces the 600 recordings down to 102 flagged samples, avoiding the need to further inspect 83% of the database. It can be observed from the plot that the threshold θ, defined in (10) can provide a reasonably acceptable cutoff value for identification of outliers which leads to detecting 97.4% of outliers and only 2 misdetections and 27 false alarms in this data set. This is beneficial when there is no prior knowledge about the number of outliers in a database.

VI. CONCLUSION

In this paper, we proposed a simple and effective method for detecting potential low-quality and inconsistent outliers in a speech database. This approach operates on the MFCC features which are known to be sensitive to changes in signal characteristics due to noise and distortion. Assuming that the majority of recordings in a speech database have roughly the same quality, and using the deterministic MCD algorithm to estimate the robust center and scatter of the observations, the potential outliers are detected based on their robust distance from the robust center of the data. We showed that a threshold equal to $\sqrt{\chi^2_{m,0.975}}$ can provide a reasonably acceptable cutoff value for detecting outliers, particularly, when there is no prior knowledge about the number of outliers. Experimental results using eight different databases show the effectiveness of the proposed method in detecting outliers in speech databases which can significantly decrease the effort required for further inspection to manually identify and remove poor-quality samples. Future work should focus on evaluating how the proposed quality control can improve the performance of individual speech-based applications. We plan to evaluate this issue on the performance of voice-based Parkinson's disease detection.

REFERENCES

[1] M. Woelfel and J. P. McDonough, *Distant speech recognition*. Wiley, 2009.

[2] K. A. Lee, H. Li, L. Deng, V. Hautamäki, W. Rao, X. Xiao, A. Larcher, H. Sun, T. Nguyen, G. Wang, A. Sizov, J. Chen, I. Kukanov, A. H. Poorjam, T. Trong, C.-L. Xu, H.-H. Xu, B. Ma, E.-S. Chng, and S. Meignier, "The 2015 NIST Language Recognition Evaluation : the Shared View of I2R, Fantastic4 and SingaMS," in *Interspeech*, San Francisco, USA, 2016, pp. 3211–3215.

[3] A. H. Poorjam, R. Saeidi, T. Kinnunen, and V. Hautamäki, "Incorporating uncertainty as a quality measure in i-vector based language recognition," in *Odyssey: the Speaker and Language Recognition Workshop*, Bilbao, Spain, 2016, pp. 74–80.

[4] A. H. Poorjam, M. H. Bahari, and H. Van hamme, "Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals," in *4th International Conference on Computer and Knowledge Engineering (ICCKE)*, 2014, pp. 7–12.

[5] S. Arora, V. Venkataraman, A. Zhan, S. Donohue, K. Biglan, E. Dorsey, and M. Little, "Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study," *Parkinsonism Related Disorders*, vol. 21, no. 6, pp. 650–653, 2015.

[6] C. Vasquez, "Automatic Detection of Parkinson's Disease from Continuous Speech Recorded in Real-World Conditions," pp. 3–7.

[7] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[8] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1992, pp. 517–520.

[9] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[10] "Tools for collecting speech corpora via mechanical-turk," in *Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010, pp. 184–187.

[11] A. Gruenstein, I. Mcgraw, and A. Sutherland, "A self-transcribing speech corpus: Collecting continuous speech with an online educational game," in *Speech, Language, and Technology in Education (SLaTE)*, 2009, pp. 109–112.

[12] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Interspeech*, Stockholm, 2017.

[13] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. Van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, J. Alam, A. Swart, and J. Perez, "The RedDots data collection for speaker recognition," in *Interspeech*, 2015, pp. 2996–3000.

[14] M. A. Little, "Parkinson's Voice Initiative." [Online]. Available: http://www.parkinsonsvoice.org/index.php

[15] 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.

[16] H. Van Den Heuvel, "Methodology for a Quick Quality Check of SLR and Phonetic Lexicons," Tech. Rep., 2004.

[17] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, 2018.

[18] M. Hubert, P. J. Rousseeuw, and T. Verdonck, “A deterministic algorithm for robust location and scatter,” *Journal of Computational and Graphical Statistics*, vol. 21, no. 3, pp. 618–637, 2012.

[19] 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.

[20] P. J. Rousseeuw, A. M. Leroy, and John Wiley Sons., *Robust regression and outlier detection*. Wiley, 1987.

[21] P. J. Rousseeuw, “Multivariate estimation with high breakdown point,” *Mathematical Statistics and Applications*, vol. 8, pp. 283–297, 1985.

[22] P. J. Rousseeuw and K. Van Driessen, “A fast algorithm for the minimum covariance determinant estimator,” *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[23] P. J. Rousseeuw and C. Croux, “Alternatives to the median absolute deviation,” *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1273–1283, dec 1993.

[24] R. A. Maronna and R. H. Zamar, “Robust estimates of location and dispersion for high-dimensional datasets,” *Technometrics*, vol. 44, no. 4, pp. 307–317, 2002.

[25] A. S. Hadi and J. S. Simonoff, “Procedures for the identification of multiple outliers in linear models,” *Journal of the American Statistical Association*, vol. 88, no. 424, p. 1264, 1993.

[26] 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.

[27] 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, 2017, pp. 289–293.

[28] P. J. Rousseeuw and B. C. van Zomeren, “Unmasking multivariate outliers and leverage points,” *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 633–639, 1990.

[29] C. Valentini-Botinhao, “Noisy speech database for training speech enhancement algorithms and TTS models,” 2016. [Online]. Available: https://datashare.is.ed.ac.uk/handle/10283/2826

[30] ——, “Noisy reverberant speech database for training speech enhancement algorithms and TTS models,” 2016. [Online]. Available: https://datashare.is.ed.ac.uk/handle/10283/2791

[31] C. Veaux, J. Yamagishi, and S. King, “The voice bank corpus: Design, collection and data analysis of a large regional accent speech database,” in *International Conference Oriental COCOSDA*, 2013, pp. 1–4.

[32] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, “The CHAINS corpus: CHAracterizing INdividual Speakers,” *International Conference on Speech and Computer (SPECOM)*, pp. 431–435, 2006.

[33] “CSTR NAM TIMIT Plus Corpus.”

[34] R. J. Moran, R. B. Reilly, P. De Chazal, and P. D. Lacy, “Telephony-based voice pathology assessment using automated speech analysis,” *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 3, pp. 468–477, 2006.

[35] G. Muhammad, M. Alsulaiman, A. Mahmood, and Z. Ali, “Automatic voice disorder classification using vowel formants,” in *2011 IEEE International Conference on Multimedia and Expo*, 2011, pp. 1–6.