Received March 9, 2020, accepted April 3, 2020, date of publication April 6, 2020, date of current version April 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2986171

Glottal Source Information for Pathological Voice Detection

N. P. NARENDRA AND PAAVO ALKU, (Fellow, IEEE)
Department of Signal Processing and Acoustics, Aalto University, Espoo 00076, Finland

Corresponding author: N. P. Narendra (narendra.prabhakera@aalto.fi)

This work was supported by the Academy of Finland under Project 312490.

ABSTRACT Automatic methods for the detection of pathological voice from healthy speech can be considered as potential clinical tools for medical treatment. This study investigates the effectiveness of glottal source information in the detection of pathological voice by comparing the classical pipeline approach to the end-to-end approach. The traditional pipeline approach consists of a feature extractor and a separate classifier. In the former, two sets of glottal features (computed using the quasi-closed phase glottal inverse filtering method) are used together with the widely used openSMILE features. Using both the glottal and openSMILE features extracted from voice utterances and the corresponding healthy/pathology labels, support vector machine (SVM) classifiers are trained. In building end-to-end systems, both raw speech signals and raw glottal flow waveforms are used to train two deep learning architectures: (1) a combination of convolutional neural network (CNN) and multilayer perceptron (MLP), and (2) a combination of CNN and long short-term memory (LSTM) network. Experiments were carried out using three publicly available databases, including dysarthric (the UA-Speech database and the TORGO database) and dysphonic voices (the UPM database). The performance analysis of the detection system based on the traditional pipeline approach showed best results when the glottal features were combined with the baseline openSMILE features. The results of the end-to-end approach indicated higher accuracies (about 2-3 % improvement in all three databases) when glottal flow was used as the raw time-domain input (87.93 % for UA-Speech, 81.12 % for TORGO and 76.66 % for UPM) compared to using raw speech waveform (85.12 % for UA-Speech, 78.83 % for TORGO and 73.71 % for UPM). The evaluation of both approaches demonstrate that automatic detection of pathological voice from healthy speech benefits from using glottal source information.

INDEX TERMS Pathological voice, glottal source waveform, glottal features, support vector machines, end-to-end systems.

I. INTRODUCTION

Pathological voice is produced as a result of different disorders affecting the human voice production mechanism [1]. Examples of voice pathologies are dysarthria [2], dysphonia [3], aphasia [4], and dyplophonia [5]. The main tasks in pathological voice processing are the detection of pathological voice from healthy speech (binary classification) [6], the detection of severity of pathology (which involves both multi-class classification and regression tasks) [7], and the identification of type of pathology (multi-class classification) [8]. From the three tasks above, the current investigation studies the first one: the detection of pathological voice from healthy speech. Pathological voice detection can be performed subjectively by speech-language pathologists which is generally regarded as a costly, laborious, and time-consuming exercise [9], [10]. Hence, an objective method for pathological voice detection can be considered as an alternative, as it is simple, reliable, and less time-consuming [11]. One of the important advantages of the objective detection is that it can be computed solely from the speech signal and performed remotely away from hospital, which helps patients to avoid frequent visits to hospital for medical examination [12]. Pathological voice detection methods can be readily integrated into on-time screening and remote health monitoring applications [13], [14].

Detection of pathological voice is an important research question despite the fact that it addresses in principle a binary

The associate editor coordinating the review of this manuscript and approving it for publication was Haiyong Zheng.

classification problem that might sound like an easy task. Voice pathology detection can be used as a diagnostic tool on its own both in voice clinics and in telemonitoring applications to screen people at risk of having a certain disease (e.g. Parkinson's disease). In addition, pathology detection methods can be used as pre-processing technologies in systems conducting more complicated tasks such as predicting the severity of pathology or recognizing a certain pathology from a group of pathologies. Even though pathological voice detection is a binary classification problem, which is in principle a more straightforward task than, for example, the prediction of the pathology severity, this research problem has certain challenges. The first challenge is the lack of pathological voice databases containing large numbers of utterances produced by large numbers of speakers. Even though there are studies (e.g. [17]) which have reported good accuracy with small amounts of speech data, the usage of fairly small databases has been shown to affect the accuracy of pathological voice detection both when using traditional classifiers and modern deep learning-based techniques [15], [16]. Compared to research areas such as speech recognition and speaker verification, where researchers have access to even hundreds of hours of training data by healthy speakers, the development of generalized and accurate detection systems remains a challenging task for pathological voice because the collection of ample amounts of data is more difficult in the domain of pathological voice. The second challenge is the variability in patient populations suffering from specific voice disorders. The patient-specific variability, which occurs across speakers and within speakers serves as a source of noise in developing pathological voice detection systems [18]. The third challenge is the presence of diverse information in voice signals (due to issues such as speaker traits, emotions, gender, and age) which complicates identifying pathological-specific factors from voice signals and makes the detection task more difficult. For interested readers, more details about the challenges in processing pathological voices can be found in a recent paper by Gupta *et al.* [18].

Based on the underlying technology, the existing studies in pathological voice detection can be broadly categorized into two approaches: *traditional pipeline* systems and modern *end-to-end* systems. In the former, the system consists of two separate parts: the feature extraction stage, called the front-end, and the classifier, called the back-end. In traditional pipeline systems, hand-crafted features extracted from speech are used to train a classifier to output labels indicating healthy/pathological. In end-to-end systems, a raw speech signal (or spectrogram) is directly used to train a deep learning model to output target labels [19], [20]. A brief review of the existing studies utilizing the two approaches in pathological voice detection is given below.

The existing works based on the traditional pipeline approach in pathological voice detection aim to combine the best speech features with the most suitable machine learning classification algorithm to effectively capture the relation between the input features and output labels. Previous studies have utilized a wide range of features presenting different aspects of speech production, such as the vocal tract (formants, line spectral frequencies (LSFs), Mel-frequency cepstral coefficients (MFCCs) and phonological features), prosody (phone duration, fundamental frequency, pitch contour, energy), voice quality (harmonic-to-noise ratio, shimmer, jitter) and glottal source (time- and frequency-domain aspects of the glottal flow) [16], [21]–[25]. For the classifier, most of the previous investigations have used support vector machines (SVMs) [6], [16], [22], [24]–[26]. In addition to SVMs, other algorithms such as artificial neural networks [22], [27], decision trees [28], linear discriminant analysis (LDA) [23], [29], and variants of recurrent neural network (RNN) [30] have also been used as classifiers in the study area. Even though existing detection studies have trained data-driven models with many different types of features, there still exists a need for novel features which are effective and robust when used with different pathological voice databases.

In studying pathological voice detection with end-to-end systems, previous studies have used either raw time-domain speech signal or its spectrum to train deep learning models [31]–[35]. In order to develop deep learning models, existing studies have mainly used combinations of convolutional neural network (CNN) and multilayer perceptron (MLP) [31], [33]–[37]. In addition, some studies have explored combining CNN and long short-term memory (LSTM) networks [32], and combining LSTM and MLP [38] for detection of pathological voice from healthy speech. Even though different deep learning architectures have been studied in the recent pathological voice detection studies listed above, a systematic comparison between latest end-to-end methods and systems based on the traditional pipeline is still lacking in the study area. In addition, it is worth noting that all the latest end-to-end systems in pathological voice detection that take advantage of raw time-domain signal waveform are based on processing the speech pressure signal. Other time-domain information signals, such as the glottal source waveform (i.e., the excitation of voiced speech, the glottal volume velocity waveform, produced by the vocal folds) are, however, possible to be utilized in training deep learning systems as shown in recent studies in text-to-speech (TTS) synthesis [39], [40]. Studying end-to-end systems that take advantage of the glottal source waveform as the raw input signal has, however, not been explored in pathological voice detection. Glottal source contains information about voice quality [41], emotion [42], and paralinguistics [43]. Compared to the speech signal, the glottal source is a more elementary time-domain signal due to the absence of vocal tract resonances. Therefore, by using this more straightforward signal as a raw signal waveform, end-to-end systems can be trained using less training data, as shown in [39].

Even though glottal flow signals have not been used as raw waveforms in training end-to-end–based pathological voice detectors, information about voice source has been utilized both in traditional pipeline systems and in end-to-end frameworks for various paralinguistic tasks such as detection of emotion, gender, dialect, and depression [44]–[48]. The importance of glottal source for pathological voice detection in the traditional pipeline framework was analyzed in [6], [25], [49], indicating discriminative power of glottal features in the identification of pathological voice. In the end-to-end system framework, the glottal source has been taken advantage of in detection of emotion [50] and depression [47], and in TTS [39]. In the area of pathological voice detection, there is typically a relatively small amount of training data available (see, e.g., [38], [51]) in comparison to speech recognition [52] and speaker verification [53] systems that might use even thousands of hours of speech data. In order to avoid the data scarcity problem, some previous studies have used data augmentation techniques to artificially increase the amount of pathological voice training data [54], [55]. However, without taking advantage of data augmentation, using deep learning models trained with voice excitation signals has shown improved accuracy in paralinguistic speech processing tasks compared to using raw speech of the same amount of data [47], [50]. Inspired by these results, the present study aims to utilize the glottal excitation for pathological voice detection in an end-to-end framework.

In this study, a comprehensive investigation of the influence of the glottal source in the detection of pathological voice is conducted by comparing traditional pipeline systems and modern end-to-end systems. In the traditional pipeline framework, acoustic features computed using the openSMILE toolkit and features representing the glottal source waveform are considered. Using both acoustic and glottal features extracted from every voice utterance and the corresponding labels indicating healthy/pathological, separate sets of SVMs are trained. In the end-to-end framework, two types of raw signal waveforms (speech and glottal source waveform estimated by glottal inverse filtering) are used to train deep learning models. Two deep learning architectures (a combination CNN and MLP as well as a combination of CNN and LSTM) are compared to build the end-to-end classifiers. Three publicly available pathological voice databases (two in dysarthria, one in dysphonia) are utilized for developing the detection systems. In the area of pathological voice detection, glottal source information has been used in the traditional pipeline framework in [6], [25], but glottal source waveforms have not been taken into account before to build end-to-end systems in the study area. Our present study is also motivated by a recent TTS study by Juvela *et al.* [39], which shows that (generative) deep learning architectures can be effectively trained using raw glottal waveforms.

The rest of the paper is arranged as follows. In section II, a brief description of pathological voice detection using a traditional pipeline framework is provided. The end-to-end system built for pathological voice detection is explained in Section III. The details about speech corpora, the experimental setup and results are provided in Section IV. Section V concludes the present study and outlines some of the possible future extensions.

II. PATHOLOGICAL VOICE DETECTION USING A TRADITIONAL PIPELINE SYSTEM

A. SYSTEM STRUCTURE

In order to classify pathological voice from healthy speech, a detection system based on the traditional pipeline approach was developed. The system, shown in Figure 1, consists of two main parts: feature extraction and SVM classifier. In the feature extraction stage, both acoustic and glottal features are computed from the speech signal. Two acoustic feature sets computed with the openSMILE toolkit are considered as baseline features. Two glottal feature sets are obtained from the glottal flow waveform, which is computed using the quasi-closed phase (QCP) algorithm [56] as the glottal inverse filtering method. Training of SVM classifiers is carried out with the acoustic and glottal features to estimate one of the two probable output labels (pathological/healthy).

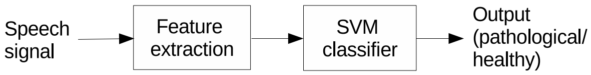

FIGURE 1. The studied pathological voice detection system based on the traditional pipeline approach.

The training phase of the pathological voice detection system is given in Figure 2. First, in order to train the detection system, three multi-speaker pathological voice databases are considered (details are provided in Section IV-A). From every speech utterance of the databases, two sets of baseline acoustic features (openSMILE-1 and openSMILE-2, described in Section II-C) are computed using the openSMILE toolkit. Glottal source waveforms are computed with the QCP method. Using the estimated glottal flows, two glottal feature sets (Glottal-1 and Glottal-2, described in Section II-B) are extracted. Using the features computed from every voice utterance as input and the corresponding output labels (pathological/healthy), training of SVM classifiers is performed. Separate SVM classifiers are trained utilizing both sets of the acoustic and glottal features, and their combinations.

After completion of the training phase, the SVM classifiers are utilized to predict the occurrence of voice pathology. During testing, the same feature sets that were used during the training phase are extracted from voice utterances. Using the extracted features as input, the SVM classifier finally outputs pathological/healthy labels.

B. GLOTTAL FEATURE EXTRACTION

In this study, the extraction of glottal features is carried out by first estimating the glottal flow from every voiced utterance using the QCP glottal inverse filtering method [56]. From the estimated flow waveforms, following two glottal feature sets are computed: time- and frequency-domain glottal features and PCA-based glottal features.

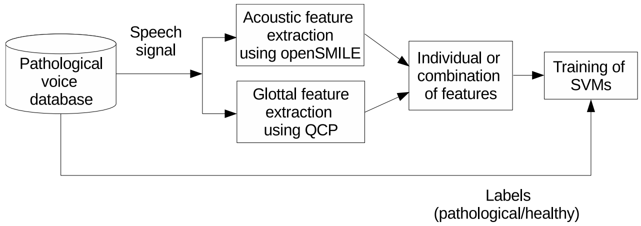
Pathological voice database
Speech signal
Acoustic feature extraction using openSMILE
Glottal feature extraction using QCP
Individual or combination of features
Training of SVMs
Labels (pathological/healthy)

FIGURE 2. Training phase of the pathological voice detection system based on the traditional pipeline approach.

TABLE 1. Glottal-1 feature set. See [59] for detailed description.

Frequency-domain glottal features	
HRF	Harmonic richness factor
H1H2	Difference between the lowest two glottal harmonics
PSP	Parabolic spectrum parameter
Time-domain glottal features	
OQ1	Open quotient, obtained from the primary glottal opening
OQ2	Open quotient, obtained from the secondary glottal opening
AQ	Amplitude quotient
OQa	Open quotient, extracted from the LF model
QOQ	Quasi-open quotient
ClQ	Closing quotient
NAQ	Normalized amplitude quotient
SQ1	Speed quotient, computed from the primary glottal opening
SQ2	Speed quotient, computed from the secondary glottal opening

1) TIME- AND FREQUENCY-DOMAIN GLOTTAL FEATURES (GLOTTAL-1)

The first glottal feature set (referred to as Glottal-1) characterizes various time- and frequency-domain aspects of the glottal flow waveform [57], [58]. A list of time- and frequency-domain glottal features used in this study is given in Table 1. Using the APARAT toolbox [59], the glottal features are derived from every cycle of the flow waveform. These features are then averaged over the frame. The glottal feature extraction is carried out in 30-ms frames only for utterances that are voiced. All features are expressed using a linear scale except for two frequency-domain glottal features (H1H2 and HRF) which are presented in the dB scale. Using the features extracted from all voiced frames of the utterance, a glottal feature vector is formed. Statistics of this glottal feature vector and its delta vector are subsequently expressed with the following eight measures: minimum, maximum, mean, median, standard deviation, range, kurtosis, and skewness. This finally leads to the Glottal-1 feature set consisting of $(12 + 12) \times 8 = 192$ features.

2) PCA-BASED GLOTTAL FEATURES (GLOTTAL-2)

The second glottal feature set (referred to as Glottal-2) parameterizes the time-domain glottal flow waveform using the principal component analysis (PCA) technique. In order to obtain PCA-based features, principal components (PCs) need to be computed first. PCs are computed using a set of glottal flow waveforms that are obtained from a speech database described in [60] consisting of long vowels. The glottal flow waveforms are differentiated, and processed in smaller segments that are two pitch periods long and centered at glottal closure indices. The segments of the flow waveform are windowed with the Hann window, interpolated to a constant length, and normalized in energy. The global mean of the flow waveform is computed using the glottal segments obtained from all utterances of the database. Every glottal segment is normalized by subtracting the global mean. Performing principal component analysis on the normalized glottal segments results in eigenvalues and eigenvectors (also known as principal components).

Using the principal components, the glottal source waveforms estimated from utterances of a pathological voice database are parameterized. By following a similar procedure as described above, two pitch period-long segments are obtained from the flow waveforms. By utilizing the principal components, each glottal segment is represented by PC weights. In this study, 30 PC weights are used to represent the glottal segments. The features obtained from all glottal segments of a frame are averaged. Using the PCA-based glottal features estimated from all voiced frames of an utterance, a glottal feature vector is formed. Following a similar procedure as in Glottal-1, the statistics related to the glottal feature vector and its delta vector are represented with eight measures, resulting in $(30 + 30) \times 8 = 480$ features representing the Glottal-2 feature set.

It is worth pointing out that both Glottal-1 and Glottal-2 represent *hand-crafted* features. Therefore, these two sets of glottal features can be used to build detectors based on the classical pipeline approach by training SVM classifiers to output pathological/healthy labels.

C. ACOUSTIC FEATURE EXTRACTION WITH openSMILE

For detecting pathological voice from healthy speech, two sets of acoustic features (referred to as openSMILE-1 and openSMILE-2) are extracted using the openSMILE toolkit [61]. In literature, features obtained with the openSMILE toolkit have been extensively utilized as baselines for the detection of paralinguistic cues such as emotions, speaker traits and states, as well as voice pathologies [62]–[64]. The first feature set consists of a set of 16 basic acoustic features given in Table 2. Using this set of acoustic features and their derivatives computed from every frame of the speech utterance, an acoustic feature vector is formed. Twelve statistical functionals (details are provided in Table 2) are computed from the acoustic feature vector to obtain $(16 + 16) \times 12 = 384$ features representing the openSMILE-1 feature set.

The second acoustic feature set includes 56 acoustic features (shown in Table 2) that are derived from every frame of the speech utterance. These acoustic features, along with their first and second order derivatives, form the acoustic feature vector. Using this acoustic vector, thirty-nine statistical functionals (details are provided in Table 2) are calculated for

TABLE 2. Two acoustic feature sets computed with the openSMILE toolkit. See [61] for detailed description.

Feature sets	Acoustic features	Statistical functionals
openSMILE-1	zero-crossing rate, RMS-energy, pitch, MFCCs (12), voicing probability	min (or max) value and its relative position, range, median, kurtosis, skewness, standard deviation, 2 linear regression coeff. and quadratic error
openSMILE-2	log-energy, zero-crossing rate, pitch, MFCCs (13), Mel-spectrum (26), jitter, shimmer, voicing probability, spectral flux, roll-off points, spectral centroid, position of spectral minimum and maximum	min (or max) value and its relative position, range, median, kurtosis, skewness, standard deviation, 2 linear regression coeff., linear and quadratic errors, 3 quartiles, 2 percentiles (95% & 98%), 3 inter-quartile errors, number of peaks, mean of peaks, mean distance between peaks, geometric, arithmetic and quadratic means

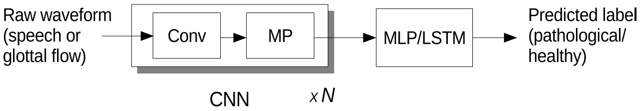
Raw waveform (speech or glottal flow)
Conv
MP
MLP/LSTM
Predicted label (pathological/ healthy)
CNN
x N

FIGURE 3. The studied pathological voice detection system based on the end-to-end approach. Conv: convolution, MP: Max pooling, MLP: multilayer perceptron, LSTM: long short-term memory, × *N*: number of convolutional layers.

every utterance to form $(56 + 56 + 56) \times 39 = 6552$ features representing the openSMILE-2 feature set.

III. PATHOLOGICAL VOICE DETECTION USING AN END-TO-END SYSTEM

The end-to-end framework for pathological voice detection is illustrated in Figure 3. The framework followed in this study has been previously used in different paralinguistic tasks [47], [48], [65]. The system consists of multiple layers of CNN followed by a multilayer perceptron. The system takes raw time-domain waveform as input, which is subsequently passed through multiple CNN layers. The resulting output from CNNs, which is a compact representation of raw waveform, is passed through MLP to predict the output label (pathological/healthy). The combination of CNN and MLP is jointly trained in a single framework. In addition to the CNN+MLP network described above, a CNN+LSTM deep learning framework is also explored in the current study.

As input to the end-to-end systems, the current study investigates two time-domain raw waveforms: the speech signal and the glottal flow waveform obtained with QCP. Instead of presenting the raw waveform over the entire voice utterance, the input is divided into fixed-length segments, which are given as inputs to the end-to-end systems. During testing, scores are obtained for every segment of an utterance. The scores obtained from all segments of the utterance are averaged and thresholded to get the final binary decision (pathological/healthy) for the utterance. In this study, 250-ms segments are used to process the raw waveform. In the development stage of the end-to-end systems, this segment duration gave the best detection accuracy compared to four other tested segment durations (50 ms, 100 ms, 500 ms and 1000 ms). This experimental result in varying the segment duration can be explained as follows. Using a too short segment duration (50 ms or 100 ms) reduces the performance of the end-to-end system as the network may not sufficiently capture pathological information present in short segments of raw waveforms. As the segment duration is increased too much (> 250 ms), the number of segments to train the end-to-end system reduces and becomes too small for the given database. Therefore, in order to properly train the detection system using a longer segment duration, more data is required. Moreover, in order to appropriately capture the embedded information in the raw waveform as the segment duration increases, the complexity of the end-to-end system grows which in turn calls for more training data. By taking into account both the amount of training data and the network complexity, it is understandable that the segment duration of 250 ms turned out to be the best choice in our search for the optimal segment duration for the end-to-end systems. The method used in this work for choosing the optimal segment duration has also been adopted in [66] and [67].

IV. EXPERIMENTS

The experiments of the current study aim to analyze the impact of glottal source information in pathological voice detection by comparing systems based on the classical pipeline approach to end-to-end systems. This section provides descriptions of the speech databases and the setup used in the experiments.

A. SPEECH CORPORA

In this study, three databases of pathological voice are used. The first two databases (UA-Speech and TORGO) consist of speech utterances collected from speakers with dysarthria, which is caused either by cerebral palsy or amyotrophic lateral sclerosis. The third corpus (the UPM database) includes utterances recorded from speakers with dysphonia, which is caused by different organic voice disorders (nodules, polyps, oedemas, and carcinomas).

1) THE UA-SPEECH DATABASE

The Universal access speech (UA-Speech) database [68] consists of speech recordings from 13 dysarthric patients (four female and nine male) and 13 healthy controls speakers (four female and nine male). The age of the subjects varies from 18 years to 58 years. A list of isolated words was uttered by every speaker in three blocks. The total number of words in each block is 255, which includes a set of 155 words that are the same in all blocks and a set of 100 words that are different across the blocks. An eight-microphone array was used to record speech utterances, and the spacing between each microphone was 1.5 inches. This work utilizes speech utterances collected from microphone no. 6 of the array with a sampling frequency of 16 kHz. Further details regarding the speech database are given in [68].

2) THE TORGO DATABASE

The TORGO database [69] contains speech utterances recorded from eight dysarthric patients (three females and five males) and seven healthy control speakers (three females and four males). The age of the speakers in TORGO varies from 16 years to 50 years. This database consists of three types of speech utterances, namely non-words, words, and sentences. Non-words include vowels uttered in high and low pitch, and multiple repetitions of /*iy-p-ah*/, /*ah-p-iy*/, and /*p-ah-t-ah-k-ah*/. Words used for speech recordings were chosen from different sources such as the word intelligibility sections of the Frenchay Dysarthria Assessment [70] and the Yorkston-Beukelman Assessment of Intelligibility of Dysarthric Speech [71]. Sentences were obtained from different sources such as the Grandfather passage from the Nemours database [72], sentence intelligibility section of the Yorkston-Beukelman Assessment of Intelligibility of Dysarthric Speech [71], the MOCHA database [73], and spontaneously elicited descriptive texts. Detailed descriptions about the text prompts and recordings are given in the TORGO database paper [69]. In this work, all three types of speech utterances that were recorded by the array microphone with a sampling frequency of 16 kHz are used.

3) THE UPM DATABASE

The Universidad Politécnica de Madrid (UPM) database [74], [75] consists of speech recordings from 201 speakers with dysphonia (75 male and 126 female) and 201 healthy control speakers (88 male and 113 female). The age of the speakers varies from 11 years to 76 years. Each speaker produced the Spanish /a/ vowel using sustained phonation. The vowels were originally sampled at 50 kHz, but the data was downsampled to 16 kHz for the purposes of the present study.

B. EXPERIMENTAL SETUP

The speech data from the three pathological voice databases are used in the experiments as follows. For every database, 70% of the data is used in training, 20% is used in testing and the remaining 10% of the speech data is used for validation. This type of data partition has been followed in several previous detection studies related both to traditional pipeline [76] and end-to-end [32], [33] systems. For UA-Speech and TORGO, the database is split in order to maintain a good partition of speakers with different severities or intelligibility scores between the training, validation, and test sets, without having any overlap in speakers between the different sets. A similar method of data splitting has been used, for example, in [38]. Tables 3 and 4 show the data partition in UA-Speech and TORGO, respectively. In both tables, F* and M* denote dysarthric female and male speakers, respectively, and FC* and MC* denote healthy female and male controls, respectively. In the UPM database, 280 randomly selected subjects (140 pathological and 140 healthy) are used for training, 80 randomly selected subjects (40 pathological and 40 healthy) are used for testing, and the remaining 42 subjects (21 pathological and 21 healthy) are used for validation.

TABLE 3. Data partition in the UA-Speech database. Intelligibility levels of dysarthric speakers are very low (VL), low (L), mediocre (M), and high (H).

Training	Testing	Validation
F03 (VL), F05 (H), M01 (VL), M07 (L), M08 (H), M11 (M), CF02, CF04, CM01, CM04, CM08, CM12	F02 (L), M05 (M), M10 (H), M16 (L), CF05, CM06, CM10, CM13	F04 (M), M04 (VL), M09 (H), CF03, CM05, CM09,

TABLE 4. Data partition in the TORGO database. Severities of dysarthric speakers are very low (VL), low (L), and mediocre (M).

Training	Testing	Validation
F01 (L), F03 (VL), M02 (M), M05 (L), FC02, MC03, MC04	F04 (VL), M04 (M), FC03, MC01	M01 (M), M03 (VL), FC01, MC02

Using the three pathological voice databases, detection systems are developed based on the traditional pipeline and end-to-end approaches. In order to develop a system based on the traditional pipeline approach, the speech utterances of the three databases are analyzed with a frame size of 30 ms in 15-ms increments. Two acoustic feature sets (openSMILE-1 and openSMILE-2) are computed from every utterance of the databases. The glottal flow waveforms are estimated from voiced frames of every speech utterance using QCP. Using the estimated glottal flow waveforms, two glottal feature sets (Glottal-1 and Glottal-2) are computed. Using global mean and global standard deviation computed from acoustic and glottal features, each of the features are individually normalized. Training of SVM classifiers is performed using the normalized acoustic and glottal features. Utilizing both individual as well as a combination of acoustic and glottal feature sets, training of separate sets of SVM classifiers is carried out. Gaussian, radial basis function kernel is used to train the SVM classifiers. The kernel parameter γ and penalty parameter C are optimally selected for every SVM classifier. The values of C and γ are varied from 10^{-3} to 10^{3} in multiples of 10 and the pair (C, γ) which leads to the highest classification accuracy with the validation data is selected.

In the end-to-end systems, speech utterances of the three databases are treated in 250-ms segments with a 50-ms increment. Glottal flow waveforms estimated with QCP are treated in similar segments. The segments of the raw speech waveform and the raw glottal flow waveform are used to separately train CNN+MLP and CNN+LSTM networks. Both CNN+MLP and CNN+LSTM networks are individually trained for each of the three databases. The details of the network architecture are given in Table 5. In this work, only 3 layers are used in CNNs in the end-to-end systems. The network configuration used in this study has been

TABLE 5. Network architecture of the end-to-end systems.

Network	Configuration
CNN+MLP	conv1: filters = 16, kernel size = 64, strides = 2, conv2: filters = 32, kernel size = 32, strides = 2, conv3: filters = 64, kernel size = 16, strides = 2, MLP: 128 hidden units, Activation: ReLu Fully connected output layer, Activation: Sigmoid
CNN+LSTM	conv: same as above LSTM: 128 hidden units Fully connected output layer, Activation: Sigmoid

successfully used in different paralinguistic recognition tasks in previous studies by other authors [47], [48], [65]. These previous studies indicate that in the context of paralinguistic recognition tasks, increasing the number of layers does not have a significant effect on the classification accuracy. Every convolution layer is followed by a ReLu activation function and MaxPooling operation with a pool size of 2, and a stride of 2. During the training stage, batch normalization is used to reduce the issue of internal covariate shift by normalizing layer inputs [77], and dropout is used to avoid over-fitting of deep neural networks [78]. The parameters of the systems are optimized using the stochastic gradient descent algorithm with the binary cross entropy-based error criterion.

In order to quantify the performance of different pathological voice detectors, the following three metrics are used: classification accuracy, sensitivity, and specificity. The classification accuracy is determined as the ratio of the number of voice utterances that are classified correctly to the total number of utterances. Sensitivity is the ratio between the number of pathological voice utterances that are correctly classified and the total number of pathological voice utterances. Specificity is the proportion of healthy speech utterances that are correctly classified.

C. RESULTS

During testing, classification accuracy, sensitivity, and specificity values of the different classifiers developed are computed. Classification results for the different feature sets of the SVM classifiers are shown in Table 6 for the three pathological voice databases. From the table, it can be observed that the classification results (accuracy, sensitivity, and specificity) obtained for the two sets of the openSMILE features are higher compared to the glottal feature sets for all three databases except for sensitivity in UA-Speech and specificity in UPM. For the two openSMILE feature sets, openSMILE-2 shows better classification results in all three measures and three databases, except for sensitivity in the UPM database. Among the three speech databases, classification accuracies are higher for the UA-Speech database compared to the TORGO and UPM databases, except for the openSMILE-1 set and the openSMILE-1 + Glottal-2 set, which gave the best accuracy in TORGO. Accuracies of the classifiers developed using the Glottal-1 and Glottal-2 features vary in the range of 63–73% for the three databases. This result shows that the glottal source consists of discriminative information needed for the detection of pathological voice from healthy speech. Importantly, by combining these glottal feature sets with the openSMILE feature sets, an improvement in classification results can be observed. This signifies the complementary nature of glottal features, which leads to improved accuracy, sensitivity, and specificity when combined with the openSMILE features. Among the different feature combinations, the openSMILE-2 + Glottal-1 feature set led to the best results for UA-Speech and TORGO, except for specificity in TORGO. For the UPM database, however, the openSMILE-2 + Glottal-2 feature set led to the best results.

The classification results of the end-to-end systems using both raw speech signals and raw glottal flow waveforms are shown for the three pathological voice databases in Table 7. From the table, it can be observed that in terms of all three measures (accuracy, sensitivity, and specificity) the CNN+MLP network performs better than the CNN+LSTM network for all three databases, except for sensitivity in the UPM database. Among the three databases, UA-Speech shows higher classification accuracies compared to the TORGO and UPM databases. In addition, the use of the raw glottal waveform shows higher values of accuracy, sensitivity, and specificity compared to using the raw speech signal for all databases and for both of the two networks except for specificity obtained with CNN-MLP in the TORGO database. This important finding indicates two issues. First, the better performance obtained by using glottal flow as raw waveform indicates that the underlying voice pathologies affect the fluctuation of the vocal folds, as reported in [79], [80]. If the voice pathologies studied affected solely some other parts of voice production (e.g., duration, vocal tract), the accuracy obtained by using speech signal as raw waveform would have been better. Second, since the glottal flow serves as the acoustical excitation of the vocal tract, glottal information is also embedded in the produced signal; that is, in the raw speech waveform. However, in addition to glottal information, raw speech waveforms also include phonemic and speaker-specific information that is brought about by the vocal tract. Involvement of phonemic and speaker-specific information makes learning of the detection problem more difficult for the deep learning networks if there is only a relatively small amount of training data available (which is the case in the current study due to traning the systems with pathological voices).

Using the data presented in Tables 6 and 7, it is possible to compare the performances of the classifiers developed with the traditional pipeline approach to the end-to-end systems. In terms of the classification accuracy, it can be noticed that the best traditional pipeline classifier, trained using the combination for the openSMILE and glottal features, is better than the best end-to-end system for all three databases. The difference between the two detection approaches is, however, small, except in the UA-Speech database where the classical pipeline system gave an absolute improvement in accuracy that was almost 4% compared to the best end-to-end system (i.e., 91.88% vs. 87.93%).

TABLE 6. Classification results obtained using the classical pipeline systems with the openSMILE and glottal features for the three pathological voice databases.

	UA-Speech			TORGO			UPM		
Feature set	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
openSMILE-1	76.65	73.76	79.54	78.24	72.94	83.54	70.73	78.49	62.98
openSMILE-2	86.99	82.35	91.63	80.62	73.73	87.52	72.50	77.61	67.84
Glottal-1	73.56	76.47	70.65	67.17	71.22	63.12	63.17	67.08	59.27
Glottal-2	68.74	69.36	68.12	66.93	71.55	62.17	64.63	65.85	63.41
openSMILE-1 + Glottal-1	81.01	82.55	79.48	79.62	73.21	86.03	73.39	79.54	67.24
openSMILE-2 + Glottal-1	**91.88**	**92.56**	**91.21**	**82.12**	**79.02**	85.22	75.61	79.11	72.20
openSMILE-1 + Glottal-2	80.02	81.93	78.12	80.63	72.59	**88.68**	74.17	80.05	68.29
openSMILE-2 + Glottal-2	91.19	91.57	90.82	81.35	76.83	85.87	**76.83**	**80.81**	**73.36**

TABLE 7. Classification results obtained using the end-to-end systems for the three pathological voice databases. The results are averaged over 5 runs and mean values are reported.

		UA-Speech			TORGO			UPM		
Network	Input	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy % (%)	Sensitivity (%)	Specificity (%)	Accuracy % (%)	Sensitivity (%)	Specificity (%)
CNN+MLP	Raw speech	85.12	80.85	89.40	78.83	82.85	**76.24**	73.71	75.92	71.35
	Glottal flow	**87.93**	**86.65**	**90.62**	**81.12**	**85.88**	75.26	**76.66**	**80.50**	**73.50**
CNN+LSTM	Raw speech	74.19	69.26	81.48	71.15	78.45	66.17	63.24	65.53	59.92
	Glottal flow	77.57	73.13	82.48	75.41	81.32	69.68	72.22	76.56	68.14

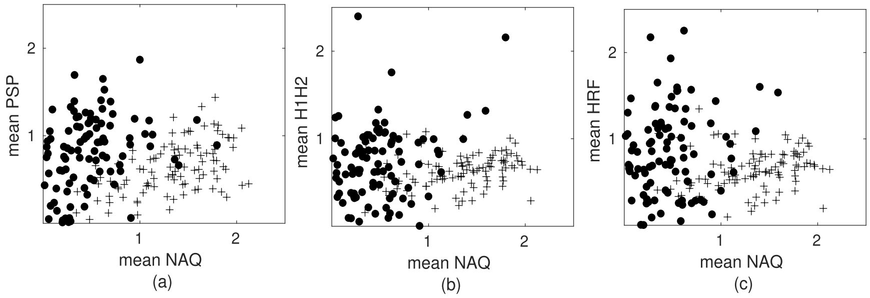

FIGURE 4. Scatter plots between pairs of glottal features for pathological and healthy voices from 100 randomly selected utterances of the TORGO database. The glottal features used are the normalized amplitude quotient (NAQ) [57], parabolic spectral parameter (PSP) [81], the difference between the amplitudes of the lowest two glottal harmonics (H1H2) [82], and harmonic richness factor (HRF) [58]. Parameters are expressed using three pairs: (a) mean NAQ vs. mean PSP, (b) mean NAQ vs. mean H1H2, and (c) mean NAQ vs. mean HRF. The filled circle denotes pathological voice and the '+' mark denotes healthy voice.

In Table 7, the results obtained from the end-to-end systems containing only 3 convolutional layers in CNNs are given. In order to better understand the impact of the number of convolutional layers of CNNs on the classification accuracy, we varied the number of convolutional layers in the end-to-end systems (the CNN+MLP architecture) from 1 to 8. The classification accuracy obtained for the raw speech input of UA-Speech was 84.16 %, 84.67 % **85.12 %**, 85.01 %, 84.84 %, 84.12 %, 83.24 %, and 82.31 % for 1, 2, 3, 4, 5, 6, 7 and 8 convolutional layers of CNNs, respectively. The best result was obtained when the end-to-end system with 3 convolutional layers in CNNs was used. The accuracy obtained by using 4 and 5 convolutional layers of CNNs was, however, close to that obtained using a 3-layer network. The results obtained by CNNs of different convolutional layer numbers were, however, all worse than results given by the best traditional pipeline systems.

Figure 4 shows scatter plots for pairs of glottal features computed from 100 pathological and healthy voice samples that were computed by randomly selecting utterances from the TORGO database. Four glottal features (NAQ, PSP, H1H2 and HRF) were computed from the flow waveforms that were obtained using the QCP inverse filtering method. Even though the scatter plots show slight overlapping between the pathological and healthy populations, the data demonstrates how glottal source information enables pathological voices to be distinguished from healthy voices.

V. CONCLUSION

Pathological voice detection systems were built based on the traditional pipeline approach and on the end-to-end approach. In the traditional pipeline approach, SVM classifiers were trained to predict pathological/healthy labels using acoustic and glottal features. Acoustic features were computed using the openSMILE toolkit, and glottal features were obtained using glottal flow waveforms estimated with the QCP glottal inverse filtering algorithm. Experimental results showed fairly good classification accuracy for the glottal features, which proves the discriminative capabilities of the glottal source. Results also indicated that the combination of the glottal features and acoustic features lead to improved classification accuracy. In the end-to-end system approach, both CNN+MLP and CNN+LSTM deep learning models were trained with raw speech and glottal flow waveforms. The results indicated that the classification accuracies are higher for the deep learning models trained with glottal flow compared to raw speech.

The current study is the first detailed investigation of the effectiveness of the glottal source in pathological voice detection comparing the traditional pipeline and end-to-end system approaches. The present work could demonstrate that the glottal source and its features contain important information required to differentiate pathological voice from healthy speech. Possible future works are as follows: In addition to dysarthria and dysphonia, the present study can be explored for other voice pathologies such as dysphasia and dyplophonia. Apart from binary classification task, the effectiveness of glottal source can be explored for the detection of different types of voice pathologies. The approach studied in this work can be utilized for the detection of paralinguistic cues such as emotional states and speaker traits.

REFERENCES

[1] A. E. Aronson and D. M. Bless, *Clinical Voice Disorders*. New York, NY, USA: Thieme Medical, 2009.

[2] J. C. Rosenbek and L. L. LaPointe, "The dysarthrias: Description, diagnosis, and treatment," in *Clinical Management of Neurogenic Communication Disorders*, D. F. Johns, Ed. Boston, MA, USA: Little, Brown and Co, 1985, pp. 97–152.

[3] M. J. Aminoff, H. H. Dedo, and K. Izdebski, "Clinical aspects of spasmodic dysphonia," *J. Neurol., Neurosurgery, Psychiatry*, vol. 41, no. 4, pp. 361–365, 1978.

[4] A. Basso and M. Macis, "Therapy efficacy in chronic aphasia," *Behavioural Neurol.*, vol. 24, no. 4, pp. 317–325, 2011.

[5] P. H. Ward, J. W. Sanders, R. Goldman, and G. P. Moore, "Diplophonia," *Ann. Otol., Rhinol., Laryngol.*, vol. 78, no. 4, pp. 771–777, 1969.

[6] N. P. Narendra and P. Alku, "Dysarthric speech classification using glottal features computed from non-words, words and sentences," in *Proc. Interspeech*, Sep. 2018, pp. 3403–3407.

[7] J. Kim, M. Nasir, R. Gupta, M. V. Segbroeck, D. Bone, M. Black, Z. I. Skordilis, Z. Yang, P. Georgiou, and S. Narayanan, "Automatic estimation of Parkinson's disease severity from diverse speech tasks," in *Proc. Interspeech*, 2015, pp. 914–918.

[8] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, K. H. Malki, and M. A. Bencherif, "An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification," *J. Voice*, vol. 31, no. 1, pp. 113.e9–113.e18, Jan. 2017.

[9] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek, "Toward phonetic intelligibility testing in dysarthria," *J. Speech Hearing Disorders*, vol. 54, no. 4, pp. 482–499, Nov. 1989.

[10] G. Van Nuffelen, C. Middag, M. De Bodt, and J. Martens, "Speech technology-based assessment of phoneme intelligibility in dysarthria," *Int. J. Lang. Commun. Disorders*, vol. 44, no. 5, pp. 716–730, Jan. 2009.

[11] J. Carmichael, "Introducing objective acoustic metrics for Frenchay Dysarthria Assessment procedure," Ph.D. dissertation, Univ. Sheffield, Sheffield, U.K., 2007.

[12] C. G. Goetz, G. T. Stebbins, D. Wolff, W. DeLeeuw, H. Bronte-Stewart, R. Elble, M. Hallett, J. Nutt, L. Ramig, T. Sanger, A. D. Wu, P. H. Kraus, L. M. Blasucci, E. A. Shamim, K. D. Sethi, J. Spielman, K. Kubota, A. S. Grove, E. Dishman, and C. B. Taylor, "Testing objective measures of motor impairment in early Parkinson's disease: Feasibility study of an at-home testing device," *Movement Disorders*, vol. 24, no. 4, pp. 551–556, Mar. 2009.

[13] P. Klumpp, T. Janu, T. Arias-Vergara, J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth, "Apkinson—A mobile monitoring solution for Parkinson's disease," in *Proc. Interspeech*, Aug. 2017, pp. 1839–1843.

[14] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1015–1022, Apr. 2009.

[15] T. Arias-Vergara, J. C. Vasquez-Correa, J. R. Orozco-Arroyave, P. Klumpp, and E. Noth, "Unobtrusive monitoring of speech impairments of Parkinson'S disease patients through mobile devices," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6004–6008.

[16] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, "Automatic intelligibility classification of sentence-level pathological speech," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 132–144, Jan. 2015.

[17] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinson's disease," in *Proc. Interspeech*, 2015, pp. 95–99.

[18] R. Gupta, T. Chaspari, J. Kim, N. Kumar, D. Bone, and S. Narayanan, "Pathological speech processing: State-of-the-art, current challenges, and future directions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 6470–6474.

[19] K. Xu, B. Zhu, Q. Kong, H. Mi, B. Ding, D. Wang, and H. Wang, "General audio tagging with ensembling convolutional neural networks and statistical features," *J. Acoust. Soc. Amer.*, vol. 145, no. 6, pp. EL521–EL527, Jun. 2019.

[20] E. Fonseca, R. Gong, and X. Serra, "A simple fusion of deep and shallow learning for acoustic scene classification," in *Proc. Sound Music Comput. Conf. (SMC)*, 2018, pp. 265–272.

[21] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Proc. Joint EMBS/BMES Conf.*, 2002, p. 182.

[22] F. Rudzicz, "Phonological features in discriminative classification of dysarthric speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2009, pp. 4605–4608.

[23] P. Gómez-Vilda, R. Fernández-Baillo, V. Rodellar-Biarge, V. N. Lluis, A. Álvarez-Marquina, L. M. Mazaira-Fernández, R. Martínez-Olalla, and J. I. Godino-Llorente, "Glottal source biometrical signature for voice pathology detection," *Speech Commun.*, vol. 51, no. 9, pp. 759–781, Sep. 2009.

[24] V. Uloza, A. Verikas, M. Bacauskiene, A. Gelzinis, R. Pribuisiene, M. Kaseta, and V. Saferis, "Categorizing normal and pathological voices: Automated and perceptual categorization," *J. Voice*, vol. 25, no. 6, pp. 700–708, Nov. 2011.

[25] N. P. Narendra and P. Alku, "Dysarthric speech classification from coded telephone speech using glottal features," *Speech Commun.*, vol. 110, pp. 47–55, Jul. 2019.

[26] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londono, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Automatic detection of Parkinson's disease from words uttered in three different languages," in *Proc. Interspeech*, 2014, pp. 1573–1577.

[27] L. A. Forero M., M. Kohler, M. M. B. R. Vellasco, and E. Cataldo, "Analysis and classification of voice pathologies using glottal signal parameters," *J. Voice*, vol. 30, no. 5, pp. 549–556, Sep. 2016.

[28] D. Hemmerling, J. R. Orozco-Arroyave, A. Skalski, J. Gajda, and E. Nöth, "Automatic detection of Parkinson's disease based on modulated vowels," in *Proc. Interspeech*, Sep. 2016, pp. 1190–1194.

[29] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of Alzheimer's disease in conversational german," in *Proc. Interspeech*, Sep. 2016, pp. 1938–1942.

[30] A. Mayle, Z. Mou, R. Bunescu, S. Mirshekarian, L. Xu, and C. Liu, "Diagnosing dysarthria with long short-term memory networks," in *Proc. Interspeech*, Sep. 2019, pp. 4514–4518.

[31] A. Rueda and S. Krishnan, "Augmenting dysphonia voice using Fourier-based synchrosqueezing transform for a CNN classifier," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6415–6419.

[32] P. Harar, J. B. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget, and Z. Smekal, "Voice pathology detection using deep learning: A preliminary study," in *Proc. Int. Conf. Workshop Bioinspired Intell. (IWOBI)*, Jul. 2017, pp. 1–4.

[33] H. Wu, J. Soraghan, A. Lowit, and G. Di-Caterina, "A deep learning method for pathological voice detection using convolutional deep belief networks," in *Proc. Interspeech*, Sep. 2018, pp. 446–450.

[34] H. Wu, J. Soraghan, A. Lowit, and G. Di Caterina, "Convolutional neural networks for pathological voice detection," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 1–4.

[35] M. Alhussein and G. Muhammad, "Voice pathology detection using deep learning on mobile healthcare framework," *IEEE Access*, vol. 6, pp. 41034–41041, 2018.

[36] H. Eghbal-zadeh, B. Lehner, M. Dorfer, and G. Widmer, "A hybrid approach with multi-channel i-vectors and convolutional neural networks for acoustic scene classification," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2017, pp. 2749–2753.

[37] R. Gong, E. Fonseca, D. Bogdanov, O. Slizovskaia, E. Gomez, and X. Serra, "Acoustic scene classification by fusing LightGBM and VGG-net multichannel predictions," in *Proc. IEEE AASP Detection Classification Acoustic Scenes Events*, 2017, pp. 1–4.

[38] J. Millet and N. Zeghidour, "Learning to detect dysarthria from raw speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5831–5835.

[39] L. Juvela, B. Bollepalli, V. Tsiaras, and P. Alku, "Glotnet—A raw waveform model for the glottal excitation in statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 6, pp. 1019–1030, Mar. 2019.

[40] L. Juvela, V. Tsiaras, B. Bollepalli, M. Airaksinen, J. Yamagishi, and P. Alku, "Speaker-independent raw waveform model for glottal excitation," in *Proc. Interspeech*, Sep. 2018, pp. 2012–2016.

[41] J. Kreiman, Y.-L. Shue, G. Chen, M. Iseli, B. R. Gerratt, J. Neubauer, and A. Alwan, "Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation," *J. Acoust. Soc. Amer.*, vol. 132, no. 4, pp. 2625–2632, Oct. 2012.

[42] M. Airas and P. Alku, "Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalised amplitude quotient," *Phonetica*, vol. 63, no. 1, pp. 26–46, 2006.

[43] E. Moore, M. A. Clements, J. W. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 1, pp. 96–107, Jan. 2008.

[44] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, and B. Yegnanarayana, "Analysis of excitation source features of speech for emotion recognition," in *Proc. Interspeech*, 2015, pp. 1324–1328.

[45] K. S. Rao and S. G. Koolagudi, "Identification of Hindi dialects and emotions using spectral and prosodic features of speech," *Systemics, Cybern. Informat.*, vol. 9, no. 4, pp. 24–33, 2011.

[46] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5200–5204.

[47] S. P. Dubagunta, B. Vlasenko, and M. M. Doss, "Learning voice source related information for depression detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6525–6529.

[48] S. H. Kabil, H. Muckenhirn, and M. M. Doss, "On learning to identify genders from raw speech signal using CNNs," in *Proc. Interspeech*, Sep. 2018, pp. 287–291.

[49] S. Gillespie, Y.-Y. Logan, E. Moore, J. Laures-Gore, S. Russell, and R. Patel, "Cross-database models for the classification of dysarthria presence," in *Proc. Interspeech*, Aug. 2017, pp. 3127–3131.

[50] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition," in *Proc. Interspeech*, Sep. 2016, pp. 3603–3607.

[51] D. Korzekwa, R. Barra-Chicote, B. Kostek, T. Drugman, and M. Lajszczak, "Interpretable deep learning model for the detection and reconstruction of dysarthric speech," in *Proc. Interspeech*, Sep. 2019, pp. 3890–3894.

[52] W.-N. Hsu, D. Harwath, and J. Glass, "Transfer learning from audio-visual grounding to speech recognition," in *Proc. Interspeech*, Sep. 2019, pp. 3242–3246.

[53] T. Kinnunen, E. Karpov, and P. Franti, "Real-time speaker identification and verification," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 1, pp. 277–288, Jan. 2006.

[54] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6009–6013.

[55] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition," in *Proc. Interspeech*, Sep. 2018, pp. 471–475.

[56] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 3, pp. 596–607, Mar. 2014.

[57] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *J. Acoust. Soc. Amer.*, vol. 112, no. 2, pp. 701–710, Aug. 2002.

[58] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Amer.*, vol. 90, no. 5, pp. 2394–2410, Nov. 1991.

[59] M. Airas, H. Pulakka, T. Bäckström, and P. Alku, "A toolkit for voice inverse filtering and parametrisation," in *Proc. Interspeech*, 2005, pp. 2145–2148.

[60] N. P. Narendra, M. Airaksinen, B. Story, and P. Alku, "Estimation of the glottal source from coded telephone speech using deep neural networks," *Speech Commun.*, vol. 106, pp. 95–104, Jan. 2019.

[61] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia (MM)*, 2013, pp. 835–838.

[62] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. Interspeech*, 2009, pp. 312–315.

[63] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2009 speaker trait challenge," in *Proc. Interspeech*, 2012, pp. 254–257.

[64] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, Parkinson's and Eating condition," in *Proc. Interspeech*, 2015, pp. 478–482.

[65] J. Wagner, D. Schiller, A. Seiderer, and E. André, "Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant?" in *Proc. Interspeech*, Sep. 2018, pp. 147–151.

[66] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," in *Proc. Interspeech*, Sep. 2019, pp. 3920–3924.

[67] D. Tang, J. Zeng, and M. Li, "An end-to-end deep learning framework for speech emotion recognition of atypical individuals," in *Proc. Interspeech*, Sep. 2018, pp. 162–166.

[68] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proc. Interspeech*, 2008, pp. 1741–1744.

[69] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Lang. Resour. Eval.*, vol. 46, no. 4, pp. 523–541, Dec. 2012.

[70] P. M. Enderby, *Frenchay Dysarthria Assessment*. San Diego, CA, USA: College Hill Press, 1983.

[71] K. M. Yorkston and D. R. Beukelman, *Assessment of Intelligibility of Dysarthric Speech*. Tigard, OR, USA: C.C. Publications, 1981.

[72] X. Menendez-Pidal, J. B. Polikoff, S. M. Peters, J. E. Leonzio, and H. T. Bunnell, "The nemours database of dysarthric speech," in *Proc. 4th Int. Conf. Spoken Lang. Process. (ICSLP)*, 1996, pp. 1962–1965.

[73] A. Wrench. (1999). *The MOCHA-TIMIT Articulatory Database*. [Online]. Available: http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html

[74] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, and G. Andrade-Miranda, "Modulation spectra morphological parameters: A new method to assess voice pathologies according to the GRBAS scale," *BioMed Res. Int.*, vol. 2015, pp. 1–13, Oct. 2015.

[75] J. D. Arias-Londoño, J. I. Godino-Llorente, M. Markaki, and Y. Stylianou, "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices," *Logopedics Phoniatrics Vocol.*, vol. 36, no. 2, pp. 60–69, Jul. 2011.

[76] C. Bhat, B. Vachhani, and S. K. Kopparapu, "Automatic assessment of dysarthria severity level using audio descriptors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5070–5074.

[77] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.

[78] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, Jan. 2014.

[79] C. A. Rosen, D. Anderson, and T. Murry, "Evaluating hoarseness: Keeping your patient's voice healthy," *Amer. Family Physician*, vol. 57, pp. 2775–2782, Jun. 1998.

[80] D. Tsuji, M. Dajer, C. Ishikawa, M. Takahashi, A. Montagnoli, and A. Hachiya, "Improvement of vocal pathologies diagnosis using high-speed videolaryngoscopy," *Int. Arch. Otorhinolaryngol.*, vol. 18, no. 3, pp. 294–302, 2014.

[81] P. Alku and E. Vilkman, "A frequency domain method for parametrization of the voice source," in *Proc. 4th Int. Conf. Spoken Lang. Process. (ICSLP)*, 1996, pp. 1569–1572.

[82] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *STL-QPSR*, vol. 36, nos. 2–3, pp. 119–156, 1995.

N. P. NARENDRA received the B.E. degree in electronics and communication engineering from the Siddaganga Institute of Technology (affiliated to VTU), India, in 2009, and the M.S. and Ph.D. degrees from IIT Kharagpur, Kharagpur, India, in 2012 and 2016, respectively. Since 2017, he has been a Postdoctoral Researcher with the Department of Signal Processing and Acoustics, Aalto University, Finland. His research interests include speech synthesis, analysis and detection of pathological speech, and paralinguistic speech processing.

PAAVO ALKU (Fellow, IEEE) received the M.Sc., Lic.Tech., and Dr.Sc.(Tech) degrees from the Helsinki University of Technology, Espoo, Finland, in 1986, 1988, and 1992, respectively. He was an Assistant Professor with the Asian Institute of Technology, Bangkok, Thailand, in 1993, and an Assistant Professor and a Professor with the University of Turku, Finland, from 1994 to 1999. He is currently a Professor of speech communication technology with Aalto University, Espoo. He has published more than 200 peer-reviewed journal articles and more than 200 peer-reviewed conference papers. His research interests include analysis and parameterization of speech production, statistical parametric speech synthesis, spectral modeling of speech, speech-based biomarking of human health, and cerebral processing of speech. He serves as an Associate Editor of J. Acoust. Soc. Am.