Speech Communication 110 (2019) 47–55

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Dysarthric speech classification from coded telephone speech using glottal features

N.P. Narendra*, Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Espoo 00076, Finland

ARTICLE INFO

Keywords:
Dysarthric speech
Glottal parameters
Glottal source estimation
Glottal inverse filtering
OpenSMILE
Support vector machines
Telemonitoring

ABSTRACT

This paper proposes a new dysarthric speech classification method from coded telephone speech using glottal features. The proposed method utilizes glottal features, which are efficiently estimated from coded telephone speech using a recently proposed deep neural net-based glottal inverse filtering method. Two sets of glottal features were considered: (1) time- and frequency-domain parameters and (2) parameters based on principal component analysis (PCA). In addition, acoustic features are extracted from coded telephone speech using the openSMILE toolkit. The proposed method utilizes both acoustic and glottal features extracted from coded speech utterances and their corresponding *dysarthric/healthy* labels to train support vector machine classifiers. Separate classifiers are trained using both individual, and the combination of glottal and acoustic features. The coded telephone speech used in the experiments is generated using the adaptive multi-rate codec, which operates in two transmission bandwidths: narrowband (300 Hz - 3.4 kHz) and wideband (50 Hz - 7 kHz). The experiments were conducted using dysarthric and healthy speech utterances of the TORGO and universal access speech (UA-Speech) databases. Classification accuracy results indicated the effectiveness of glottal features in the identification of dysarthria from coded telephone speech. The results also showed that the glottal features in combination with the openSMILE-based acoustic features resulted in improved classification accuracies, which validate the complementary nature of glottal features. The proposed dysarthric speech classification method can potentially be employed in telemonitoring application for identifying the presence of dysarthria from coded telephone speech.

1. Introduction

Dysarthria is a neuro-motor disorder resulting in neurological damage of the motor component of speech production (Doyle et al., 1997). Dysarthria is generally a result of either a neurological injury (i.e., cerebral palsy, brain tumor, brain injury, stroke) or a symptom of a neurodegenerative disease (i.e., Parkinsons's disease, amyotrophic lateral sclerosis, Huntington's disease). Dysarthric speech is often associated with reduced vocal tract volume and tongue flexibility, atypical speech prosody, imprecise articulation, and variable speech rate - factors that all reduce speech intelligibility (Duffy, 2012). The assessment of speech is essential in distinguishing dysarthria from healthy speech. The speech assessment can be performed using a traditional approach, which involves speech-language pathologists performing intelligibility tests to judge the presence of dysarthria, as well as to characterize its severity (Kent, 1992). Subjective intelligibility tests are, however, costly, laborious, and frequently prone to intrinsic biases of pathologists due to familiarity with patients and their speech disorders (De Bodt et al., 2002; Van Nuffelen et al., 2009). This motivates the design of an objective method for the assessment of dysarthric speech. The assessment of dysarthric speech is carried out for two tasks: (1) to identify the presence of dysarthria from a given speech signal and (2) to estimate the severity of dysarthria. Both of these tasks are crucial diagnostic steps which help to take clinical decisions regarding the course of therapy or medication of patients. This work focuses on the former task, i.e., the identification of the presence of dysarthria.

Objective assessment for the identification of the presence of dysarthria is usually done by a data-driven model, trained on collected speech data and labels obtained from the speech-language pathologist. Objective speech-based assessment is economical and reliable, and it can be used to perform the diagnosis on a regular basis (Constantinescu et al., 2010). As speech-based diagnosis can be performed remotely, away from the hospital, it can in principle be conducted using a telemonitoring application (Ramezani et al., 2017). In order to use speech-based diagnosis systems in telemonitoring applications, the system should be capable of handling varying speech degradation conditions, for example, coding (i.e., generation of quantization noise), transmission errors, band-pass filtering, and varying background environments.

* Corresponding author.
E-mail addresses: narendra.prabhakera@aalto.fi (N.P. Narendra), paavo.alku@aalto.fi (P. Alku).

https://doi.org/10.1016/j.specom.2019.04.003
Received 22 August 2018; Received in revised form 6 March 2019; Accepted 1 April 2019
Available online 8 April 2019

In the literature, there are only a few studies on the speech-based telemonitoring of neuro-motor disorders and there are no reported works specific to telemonitoring of dysarthric speech. Little et al. (2009) proposed a new measure of dysphonia - pitch period entropy (PPE), which is robust to various noisy environments for the telemonitoring of Parkinson's disease. In Tsanas et al. (2010), clinically useful features such as recurrence period density entropy (RPDE), detrended fluctuation analysis (DFA), and PPE are explored for telemonitoring of Parkinson's disease progression. In Mandal and Sairam (2013), Klumpp et al. (2017), a general telemonitoring framework is proposed which involves regular collection of speech samples from subjects and monitoring of Parkinson's disease. Sakar et al. (2017) explored vocal features such as jitter, shimmer, harmonics-to-noise ratio, recurrence period density entropy and pitch period entropy in discriminating Parkinson's disease patients with early signs of speech disorders and healthy subjects. In Arias-Vergara et al. (2018), monitoring of Parkinson's disease progression was performed by using individual speaker models which are developed using the classical GMM-UBM technique and the i-vector approach. Except for a few recent investigations (Arias-Vergara et al., 2018; Vásquez-Correa et al., 2015), most of the previous studies on speech-based telemonitoring consider, however, only clean speech recorded under ideal conditions (i.e., speech is considered to be free from distortion such as quantization noise or environmental noise). Therefore, in order to develop an effective telemonitoring system, dysarthric speech classification needs to be studied in more realistic scenarios. In the current article, a dysarthric speech classification system that works with coded telephone speech - the data that is best suitable for telemonitoring applications - is developed.

The existing dysarthric speech classification systems extract high-dimensional acoustic features to capture the wide variabilities of sources and patterns in pathological speech. Previous works have explored a range of features including spectral features (e.g., line spectral frequencies (LSFs), Mel-frequency cepstral coefficients (MFCCs), formants), prosody features (e.g., fundamental frequency, pitch contour, phone duration, energy), voice quality features (e.g., jitter, shimmer, harmonics-to-noise ratio), perceptual features, and phonological features (Dibazar et al., 2002; Falk et al., 2012; Kim et al., 2015; Rudzicz, 2009). Only a few studies have, however, utilized *glottal features* (i.e., parameters describing the source of voiced speech, the glottal flow, generated by the vocal folds) for detecting the presence of dysarthria. In Gillespie et al. (2017), glottal parameters are utilized in combination with spectral and prosodic parameters to develop cross-database models (training on one database and testing on another database) for the identification of dysarthria. In a recent study by the present authors (Narendra and Alku, 2018), the effectiveness of glottal parameters computed using an efficient glottal inverse filtering (GIF) method, quasi-closed phase analysis (QCP) (Airaksinen et al., 2014), is demonstrated for dysarthric speech classification from three speech signal categories (non-words, words, and sentences). It is worth noting that in these previous dysarthria classification studies, both the widely used acoustic measures (i.e., spectral, prosodic, and voice quality) and glottal parameters have been computed from clean speech which is free from external degradations.

For effective telemonitoring applications, the classification task needs to be performed in realistic scenarios where speech is degraded due to, for example, low bit-rate speech coding. While there exist previous works on different speech processing tasks such as speech recognition (Medennikov et al., 2016) and speaker verification (Lei and Lopez-Gonzalo, 2009; Gallardo et al., 2014) from coded telephone speech, dysarthric speech classification has not been explored before from coded telephone speech. Even though widely used acoustic features, such as LSFs, MFCCs, and linear frequency cepstral coefficients (LFCCs), have been successfully utilized for different speech processing tasks under the coded condition (Medennikov et al., 2016; Lei and Lopez-Gonzalo, 2009; Gallardo et al., 2014), the glottal parameters have not been explored from coded telephone speech due to the known strict quality requirements of GIF analysis (Alku, 2011). In Narendra et al. (2018), illustration of degradation of glottal flows obtained from coded speech using existing GIF methods is provided. Apart from coding, existing GIF analysis is highly sensitive to non-ideal recording conditions such as ambient noise, low-frequency bias due to breath burst on the microphone, phase distortions due to the recording equipment, improper A/D conversion, and conducting GIF in such circumstances can lead to significant distortions in glottal flow estimates (Wong et al., 1979; Cinnéide et al., 2010; Childers et al., 1983). Accurate estimation of glottal flow under non-ideal condition (i.e., coding) is highly necessary and has potential in improvement of performance of dysarthric speech classification. Additionally, robust estimation of glottal parameters from coded telephone speech can not only be useful for dysarthric speech classification but also in speech-based telemonitoring of different neuro-motor disorders, speaker traits, and emotions. In the current study, the dysarthric speech classification is performed from coded telephone speech using two bandwidths, narrowband (300 Hz - 3.4 kHz) and wideband (50 Hz - 7 kHz), that have been standardized in speech transmission (3GPP TS 26.090, 2011; Järvinen, 2000).

Feature extraction techniques used for dysarthric speech classification from clean speech cannot be directly applied for coded speech. The performance of feature extraction techniques, particularly for glottal parameters, will degrade for the coded speech due to the presence of amplitude and phase distortions in the signal (Narendra et al., 2017). Conventional GIF methods, which estimate the glottal flow directly by filtering the input speech utterance with the inverse of the vocal tract model are known to be highly sensitive to even slight distortions in the input speech signal (Narendra et al., 2017; Drugman et al., 2012a; Airaksinen et al., 2015). In order to effectively estimate the glottal source from coded speech, deep neural net-based glottal inverse filtering (DNN-GIF) was recently proposed (Narendra et al., 2018; 2017). In DNN-GIF, a deep neural net (DNN), trained in a supervised manner using glottal flows estimated from clean speech, maps acoustical features (e.g., LSFs, MFCCs) of the input signal directly to a time-domain glottal flow signal. DNN-GIF is shown in Narendra et al. (2018, 2017) to estimate glottal flows accurately from coded speech, compared to state-of-the-art GIF methods. The glottal parameters, which are obtained from the flow waveform estimated using DNN-GIF can be explored for dysarthric speech classification. In addition, investigation of the complementary nature of glottal parameters when used in combination with existing acoustic features can also be performed.

This study proposes a new dysarthric speech classification method, which works for coded telephone speech and therefore suits telemonitoring applications. The study particularly explores the effectiveness of glottal parameters extracted from coded telephone speech for dysarthric speech classification. The glottal parameters are extracted from the voice source signals estimated using the recently proposed DNN-GIF method. Two sets of glottal parameters are considered: (1) time- and frequency-domain parameters, and (2) principal component analysis (PCA)-based parameters. Acoustic features extracted with the openSMILE toolkit (Eyben et al., 2013) are used as the baseline. The openSMILE-based features have previously been widely used for different paralinguistic challenges, such as the detection of speech disorders, speaker effects (such as stress, styling, depression), and emotion recognition. Using the features extracted from every speech utterance as well as its corresponding label indicating, *dysarthric/healthy*, a support vector machine (SVM) classifier is trained. Experiments are conducted using two freely available databases (Rudzicz et al., 2012; Kim et al., 2008) to systematically study the effectiveness of glottal parameters when used individually and combined with the baseline openSMILE features in the classification of dysarthric speech under the coded condition.

This paper is organized as follows. In Section 2, a detailed description about the proposed dysarthric speech classification method is provided. The details about the dysarthric speech databases, experimental setup and results are given in Section 3. Conclusions of the present study and possible future extensions are presented in Section 4.

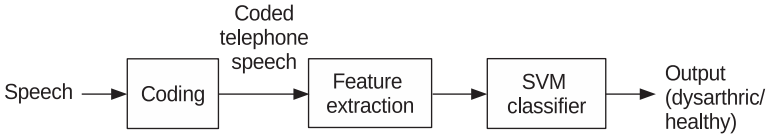

Fig. 1. The proposed dysarthric speech classification method.

2. The proposed method

2.1. System structure

In order to classify dysarthric voices from healthy speech, a speech classification system shown in Fig. 1 was developed. In this approach, SVM classifiers consider individual or combined (acoustic and glottal) features, extracted from the input speech signal (coded telephone speech), to predict one of two possible output classes (dysarthric or healthy). SVMs are widely used in pathological speech classification and they have been validated with consistent performance, even for a small amount of speech data, in contrast to other techniques such as deep neural nets which require a large amount of data for proper training (Kim et al., 2015; Orozco-Arroyave et al., 2014). In the training phase (shown in Fig. 2), the proposed method utilizes both the coded input speech signal and corresponding *dysarthric/healthy* labels obtained from speech-language pathologists. Prior to the training of the SVMs, both acoustic and glottal parameters are extracted by using suitable methods. After the SVMs have been trained, the system maps the input (a set of acoustic and glottal features extracted from coded telephone speech) to the desired output class (*dysarthric/healthy*). In this work, two sets of acoustic features, which are extracted from coded telephone speech using the openSMILE toolkit (Eyben et al., 2013) are used as reference features. Further, two sets of glottal parameters are extracted from glottal flow waveforms, which are estimated using two GIF methods (QCP and DNN-GIF) and, hence, a total of four types of glottal parameters are extracted.

In order to train the proposed dysarthric speech classifier, first, a multi-speaker dysarthric speech database is considered (described in detail in Section 3.2). The speech utterances present in the dysarthric speech databases are appropriately coded (explained in Section 2.2). From every coded speech utterance, the glottal flow waveform is estimated using two GIF methods: QCP and DNN-GIF. QCP was shown in Airaksinen et al. (2014) to be the best performing GIF method compared to existing methods (Alku, 1992; Wong et al., 1979; Drugman et al., 2009) in estimating the glottal flow from clean speech. DNN-GIF is a recently proposed method, which was shown to be the best performing GIF method under the coded condition (Narendra et al., 2018, 2017). Using the glottal flow waveform estimated from each of the two GIF methods, two sets of glottal parameters are extracted (details are provided in Section 2.3). Two sets of acoustic features, named openSMILE-1 and openSMILE-2, are also extracted from every coded speech utterance using openSMILE (described in Section 2.4), which is a widely used toolkit in paralinguistic speech processing tasks. Generally, the size of acoustic and glottal feature sets is large. To avoid the risk of over-fitting, the size of each of these feature sets is reduced by using the sequential forward feature selection (SFFS) algorithm (Reunanen, 2003). SFFS selects a subset of features from the feature set that results in the best classification accuracy. Starting from an empty feature set, SFFS creates candidate feature subsets by sequentially adding each of the features, and each candidate feature subset is evaluated by computing the classification accuracy with the 10-fold cross-validation strategy. Feature selection using SFFS results in both improvement of the computational efficiency and generalization capabilities. Using the features extracted from every coded speech utterance as input and corresponding *dysarthric/healthy* labels as output, a SVM classifier is trained. Separate classifiers are trained using reduced and non-reduced set of features for openSMILE, glottal features, and combination of these features.

After training of the SVMs is complete, the SVM classifiers can be used to identify the presence of dysarthria from the coded speech utterance. The same set of speech features, which were used during training are extracted from the coded speech utterance, and the extracted features are fed to the SVM classifier, which outputs the *dysarthric/healthy* labels.

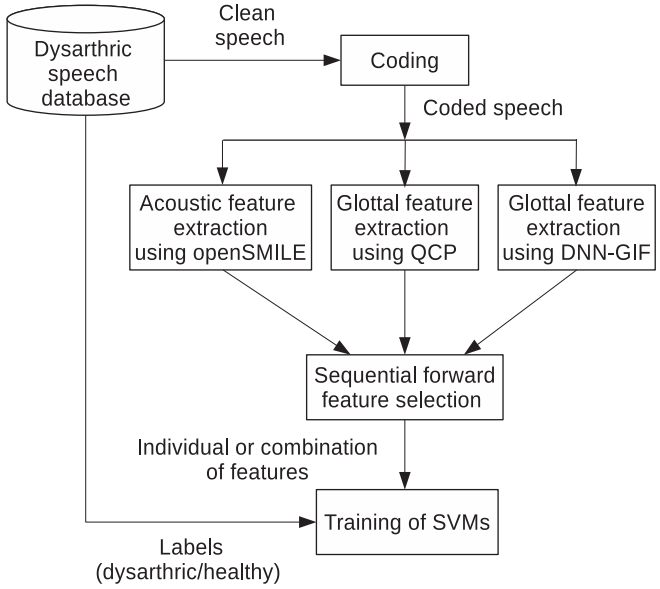

Fig. 2. The training phase of the proposed dysarthric speech classification method.

2.2. Telephone speech codecs

In order to simulate telephone speech, the adaptive multi-rate (AMR) codec (3GPP TS 26.090, 2011) is used for coding the utterances of the dysarthric speech database. The AMR codec is a widely used speech compression method, standardized by the European Telecommunications Standards Institute (ETSI) (Järvinen, 2000). This work considers AMR codecs, which operate on two transmission bandwidths - the narrowband (300 Hz - 3.4 kHz) and the wideband (50 Hz - 7 kHz). Depending on the transmission bandwidths of operation, the AMR can be categorized as either the AMR narrowband (NB) codec or the AMR wideband (WB) codec. The AMR-NB and AMR-WB codecs use sampling frequencies of 8 kHz and 16 kHz respectively. In the proposed classification system, both acoustic and glottal features are extracted separately from both NB- and WB-coded speech. Separate SVM classifiers (one for the NB-coded speech, the other for the WB-coded speech) are trained using the features extracted from AMR-coded speech as well as labels indicating, *dysarthric/healthy*.

2.3. Glottal parameter extraction

In dysarthria, as the motor component of speech production is affected, the nature of vibration of the vocal folds is deviated compared to healthy speech. The difference in vocal fold vibration between dysarthric and healthy speech production cannot be characterized completely by the *rate* of vibration (i.e., pitch information). Instead, the *mode*

of vibration of vocal folds needs to be taken into account as well. Therefore, the waveform of the acoustic speech excitation generated by the vocal folds, the glottal flow, may have useful discriminating information for dysarthric speech classification. In order to parameterize the glottal source, the flow waveform must be estimated first with GIF from coded speech signal. In order to estimate the glottal flow from coded speech, two GIF methods are utilized: QCP and the recently proposed DNN-GIF method. Here, it is worth noting that the glottal parameters are extracted only in those regions of coded speech where the vibration of the vocal folds takes place (i.e. in voiced segments).

2.3.1. GIF methods

QCP (Airaksinen et al., 2014) is one of the most accurate GIF methods used for estimating the glottal flow from clean speech. The QCP method is based on the principles of closed phase analysis (CP) (Wong et al., 1979), which estimates the vocal tract response using the covariance method of linear prediction from a few speech samples located in the closed phase of glottal cycle. In contrast to the CP method, QCP creates a specific temporal weighting function, called the attenuated main excitation (AME) function (Alku et al., 2013), using glottal closure instants (GCIs) estimated from speech. The AME function is used to attenuate the contribution of the (quasi-) open phase in the computation of the weighted linear prediction (WLP) coefficients, which results in good estimates of the vocal tract transfer function. The evaluation results in Airaksinen et al. (2014) show that the accuracy of QCP is better than that of CP (Wong et al., 1979), iterative adaptive inverse filtering (IAIF) (Alku, 1992), and complex cepstral decomposition (CCD) (Drugman et al., 2009). Even though it is shown in Narendra et al. (2018) that the glottal flow waveforms estimated with conventional GIF methods, including QCP, are distorted due to coding, the main reason for using QCP in this study is to understand how much the coding of speech affects dysarthric speech classification accuracy when glottal features are computed with conventional GIF methods.

DNN-GIF (Narendra et al., 2018, 2017) is a recently proposed GIF method to estimate the glottal source from coded speech. DNN-GIF is a data-driven method, which utilizes both coded and clean versions of speech signals during the training phase. Unlike the existing GIF methods, which attempt to accurately model the vocal tract filter and estimate the glottal flow by removing the contribution of the vocal tract from speech, the DNN-GIF method relies on the non-linear mapping capability of DNNs to estimate the glottal flow. In this method, a DNN is trained to map the spectral features (e.g., LSFs) extracted from coded telephone speech with the corresponding reference glottal flow waveforms obtained from clean speech. The reference glottal flow waveforms are estimated from clean speech by using the QCP method. During the testing phase, the spectral features extracted from coded telephone speech are given as input to the trained DNN which generates the glottal flow waveform. In DNN-GIF, separate deep neural networks are trained for NB- and WB-coded speech. According to the experiments carried out in Narendra et al. (2018, 2017), the accuracy of DNN-GIF is better than that of QCP (Airaksinen et al., 2014), CP (Wong et al., 1979), IAIF (Alku, 1992), and CCD (Drugman et al., 2009). In the current study, we take advantage of the same DNN which was used in Narendra et al. (2018, 2017) and which was trained using speech samples (both NB- and WB-coded) of sustained long vowels. The DNN trained in Narendra et al. (2018, 2017) can be used also in the current study because the glottal flow waveform to be estimated is a simple and elementary waveform generated at the level of the vocal folds (i.e., in the absence of vocal tract resonances) and mismatch between training and test data hardly have any influence on the glottal flow estimation.

Fig. 3 illustrates the glottal flow waveforms estimated from clean and coded speech by using the QCP and DNN-GIF methods. Clean speech is obtained from the vowel section /a/ of a dysarthric speech utterance of the TORGO database and the AMR-NB codec is used to obtain the corresponding coded telephone speech version. The figure shows the glottal flow waveform estimated from clean speech by using the QCP method

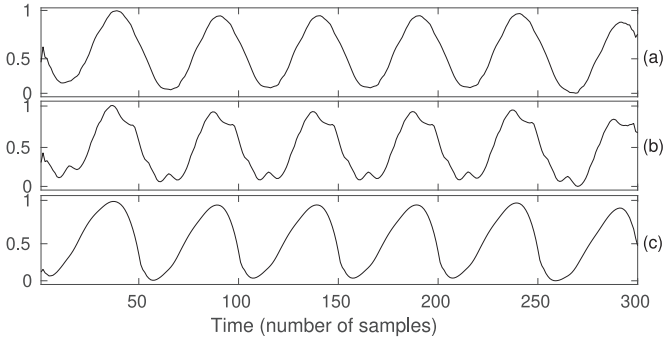

Fig. 3. (a) Glottal flow waveform obtained from clean speech by using the QCP method. Glottal flow waveforms estimated from coded speech by using (b) the QCP method and (c) the DNN-GIF method.

(Fig. 3(a)), and the glottal flow waveforms estimated from the corresponding coded speech signal by using the QCP method (Fig. 3(b)) and the DNN-GIF method (Fig. 3(c)). From the figure, it can be observed that the glottal flow waveform estimated from coded telephone speech by using the QCP method is distorted, and the glottal flow waveform estimated from coded telephone speech by using the DNN-GIF method is clearly closer to the glottal flow waveform estimated from clean speech.

Using the glottal flow waveforms estimated by GIF, two sets of glottal parameters are extracted in the current study. The first parameter set captures the time- and frequency-domain characteristics of glottal flow waveform. The second parameter set aims to represent the entire glottal flow waveform using PCA.

2.3.2. Time- and frequency-domain glottal parameters (Glottal-1)

The first glottal parameter set (referred in this work as Glottal-1) consists of 12 known time- and frequency-domain parameters, which characterize various aspects of the glottal flow waveform (Alku et al., 2002; Childers and Lee, 1991). These parameters are extracted using APARAT Toolbox (Airas et al., 2005) and they are listed in Table 1. H1H2 and HRF are obtained in the dB scale, and the other parameters are obtained in a linear scale. The glottal parameters are computed in 30-ms frames. H1H2 and HRF are computed pitch-asynchronously once per frame, whereas the rest of the parameters are computed pitch-synchronously once per glottal cycle and then averaged over the frame. The glottal parameters computed from all voiced frames of the input coded speech signal finally form the glottal parameter vector of the utterance. The following 8 statistical measures are computed from the

Table 1
Time- and frequency-domain glottal parameters. For more details, see Airas et al. (2005).

Time-domain parameters	
OQ1	Open quotient, calculated from the primary glottal opening
OQ2	Open quotient, calculated from the secondary glottal opening
NAQ	Normalized amplitude quotient
AQ	Amplitude quotient
ClQ	Closing quotient
OQa	Open quotient, derived from the LF model
QOQ	Quasi-open quotient
SQ1	Speed quotient, calculated from the primary glottal opening
SQ2	Speed quotient, calculated from the secondary glottal opening
Frequency-domain parameters	
H1H2	Difference between first two glottal harmonics
PSP	Parabolic spectral parameter
HRF	Harmonic richness factor

Table 2
Two openSMILE feature sets. For more details, see Eyben et al. (2013).

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

glottal parameter vector, as well as from its delta vector: mean, median, minimum, maximum, standard deviation, range, skewness, and kurtosis. This results in $(12 + 12) \times 8 = 192$ parameters representing the Glottal-1 parameter set.

2.3.3. PCA-based glottal parameters (Glottal-2)

The second glottal parameter set (referred in this work as Glottal-2) consists of parameters that represent the entire glottal flow waveform using PCA, which uses an orthogonal transformation to convert a set of observations into a set of linearly uncorrelated variables called principal components (PCs). The use of PCA for voice source waveform modeling was first proposed in Thomas et al. (2009). The parameterization of glottal flow using PCA has been explored for speech synthesis (Raitio et al., 2013) and speaker recognition (Drugman and Dutoit, 2012). However, PCA-based glottal flow parameterization has not been explored before for pathological speech classification.

In order to estimate the PCA-based parameters from the glottal flow waveforms of the dysarthric speech database, first, the principal components are obtained from the glottal flow waveforms that are estimated using the clean speech database containing sustained long vowels (the same speech data, which was used for training in DNN-GIF). The details of the speech database are provided in Narendra et al. (2018, 2017). The glottal flow waveforms are estimated using QCP. In order to perform PCA, the glottal flow waveform of every utterance is pitch-synchronously decomposed into smaller segments. The glottal segments are computed from the derivative of the glottal flow, which is decomposed as GCI-centered, two-pitch-period-long segments. The glottal segments are then windowed with the Hann window, interpolated to a constant length, and normalized in energy. The glottal segments are extracted from the utterances of the entire speech database and each of the glottal segments is normalized by subtracting the global mean of the glottal flow. The principal component analysis is performed on the normalized glottal segments, resulting in the computation of eigenvalues and eigenvectors (also called the principal components).

Using the principal components, the parameterization of glottal flow waveforms is performed. First, the glottal flow waveforms are estimated from the coded speech utterances of the dysarthric speech database using QCP and DNN-GIF. Using the same procedure as explained above, the glottal flow waveforms are decomposed into two-pitch-period-long glottal flow segments. The glottal segments are parameterized by projecting them on an orthonormal basis obtained by the principal components leading to PC weights. In this work, the glottal flow waveform is parameterized using 30 PC weights. The glottal parameters represented as 30 PC weights are computed from every cycle of the glottal flow and then averaged over the frame. PCA-based glottal parameters, computed from all the voiced frames of the coded speech utterance, finally form the glottal parameter vector. 8 statistical measures are computed from the vector of glottal parameters, as well as from its delta, resulting in $(30 + 30) \times 8 = 480$ parameters representing the Glottal-2 parameter set.

2.4. Parameter extraction with openSMILE

Acoustic parameters are extracted from coded speech using openSMILE, a freely available feature extraction toolkit (Eyben et al., 2013). The openSMILE features have been used as baselines for different paralinguistic challenges from INTERSPEECH 2009 (Schuller et al., 2009). Some examples of paralinguistic challenges are the recognition of emotion, speaker traits and states, and speech pathology. The acoustic features extracted by openSMILE mainly represent spectrum, prosody, and voice quality. In this work, two sets of acoustic features, defined in the openSMILE toolkit, are used for dysarthric speech classification. The first set (referred in this work as openSMILE-1) is INTERSPEECH 2009 Emotion Challenge (Schuller et al., 2009) feature set, consisting of 384 features. This feature set consists of the basic acoustic features such as MFCCs, root mean square (RMS) energy, pitch, zero-crossing rate, and voicing probability. A set of 16 acoustic features extracted from every frame is described in Table 2. The set of 16 acoustic features, along with their derivatives obtained from all frames of a speech utterance, forms the acoustic feature vector. 12 statistical functionals (shown in Table 2) are computed from the acoustic feature vector of the utterance to obtain $(16 + 16) \times 12 = 384$ features representing the openSMILE-1 feature set.

The second set (referred in this work as openSMILE-2) is the large openSMILE emotion feature set consisting of 6552 features. This is the largest feature set in terms of the number of features in the openSMILE toolkit. The largest feature set is chosen to involve as much acoustic information as possible, which may be helpful in dysarthric speech classification. A set of 56 acoustic features (given in Table 2) are extracted from every frame. 56 acoustic features, along with their first and second order derivatives, form the frame-level acoustic features. As in openSMILE-1, statistical functionals are applied on the acoustic feature vectors, which are extracted from all the frames of the speech utterance. Instead of 12, 39 statistical functionals (shown in Table 2) are applied to obtain $(56 + 56 + 56) \times 39 = 6552$ features representing the openSMILE-2 feature set. The set of acoustic features and statistical functionals used in the openSMILE-2 feature set are described in Table 2.

3. Experiments

The experiments conducted in this study evaluate the effectiveness of the glottal parameters obtained from the QCP and DNN-GIF methods in dysarthric speech classification under the coded condition. Classification accuracies obtained with different combinations of the glottal and openSMILE features are analyzed. Differences between classifiers obtained from the baseline openSMILE features and those obtained from the combination of glottal and baseline openSMILE features are also analyzed using statistical tests. To develop dysarthric speech classification systems, the TORGO (Rudzicz et al., 2012) database and the universal access speech (UA-Speech) (Kim et al., 2008) database are utilized.

3.1. The TORGO speech database

The TORGO database contains speech recordings from seven patients (three females and four males), diagnosed with cerebral palsy or amyotrophic lateral sclerosis (ALS), and speech recordings from seven healthy control speakers (three females and four males). The age range of the patients is from 16 years to 50 years. The patients were known to have disruptions in the neuro-motor interface, which causes dysarthria. The database includes speech signals in three categories, namely non-words, words, and sentences. The non-words consist of 5–10 repetitions of $/iy - p - ah/$, $/ah - p - iy/$, and $/p - ah - t - ah - k - ah/$ and high-pitched and low-pitched vowels, maintained over 5 s (e.g., "Say 'eee' in a high pitch for 5 s"). The text prompts used to record short words include 50 words from the word intelligibility section of the Frenchay Dysarthria Assessment (Enderby, 1983) and 360 words from the word intelligibility section of the Yorkston-Beukelman Assessment of Intelligibility of Dysarthric Speech (Yorkston and Beukelman, 1981). The sentences comprise three preselected phoneme-rich sentence sets: the Grandfather passage from the Nemours database (Menendez-Pidal et al., 1996), 162 sentences from the sentence intelligibility section of the Yorkston-Beukelman Assessment of Intelligibility of Dysarthric Speech (Yorkston and Beukelman, 1981), 460 sentences from the MOCHA database (Wrench, 1999), and spontaneously elicited descriptive texts. Details of text prompts and the reason for the selection of these prompts are described in the TORGO database paper (Rudzicz et al., 2012). The speech data of TORGO was recorded simultaneously through two microphones - a head-mounted microphone and an array microphone, with a 16 kHz sampling frequency. In this study, speech samples recorded by the array microphone are used. 80 sentence-level utterances from each speaker are used (except for two dysarthric speakers, only 23 and 28 utterances are used due to the lack of availability of recordings) to develop dysarthric speech classification systems. Dysarthria and healthy control (or non-dysarthria) labels associated with every speaker are obtained from the TORGO database.

3.2. The UA-speech database

The UA-Speech database (Kim et al., 2008) contains speech collected from 15 patients (four female and eleven male) diagnosed with cerebral palsy and speech recordings from 13 healthy control speakers (four females and nine males). The age range of the patients is from 18 to 58. Every subject is asked to utter 765 isolated word utterances in three blocks (B1, B2, and B3) and each block contains 255 words, which includes 155 words which are common to all three blocks and 100 uncommon words that differ across the blocks. The 155 words consist of 19 common computer commands (e.g., 'enter', 'tab'), 10 digits (0 to 9), 26 radio alphabet letters (e.g., 'Alpha', 'Bravo'), and the 100 most common words in the Brown corpus of written English (e.g., 'to', 'and'). One hundred uncommon words in each block were selected from children's novels digitized by Project Gutenberg (Kim et al., 2008). Dysarthric speech data was recorded using an eight-microphone array, sampled at 16 kHz and each microphone was spaced at intervals of 1.5 in. In this study, speech samples recorded from microphone no. 6 of the array were used. The data studied includes speech from 13 patients and 13 healthy controls. Speech utterances from a single block (B1) of each speaker were used. Same as in TORGO database, dysarthria and healthy control (or non-dysarthria) labels associated with every speaker are obtained from the UA-Speech database.

In context of neuro-motor disorders, it is worth emphasising that classification systems have been developed mainly using a relatively small number of speech samples recorded from patients (Falk et al., 2012; Kim et al., 2015; Little et al., 2009). This is different from other speech processing tasks, such as speech synthesis and speech recognition, where a large number of speech samples (e.g. thousands of sentences per speaker (Airaksinen et al., 2018)) is typically recorded from healthy speakers. Collecting a large number of utterances from patients is, however, challenging if not impossible. Patients feel difficulty in speaking for longer duration, particularly when the neuro-motor disorder is severe. Therefore, most of the existing freely available dysarthric speech databases (Rudzicz et al., 2012; Kim et al., 2008; Menendez-Pidal et al., 1996; Deller et al., 1993) contain speech data recorded from a small number of patients.

3.3. Experimental setup

Speech utterances from TORGO and UA-Speech are used in the experiments as follows. 10% of every speaker's utterances are used as validation data and the remaining 90% are used as training data. The training data is used to develop the SVM models and to compute the classification accuracy. The validation data is used only for feature selection and tuning of the hyper parameters. As there is no overlap between the training and validation data, the models developed will generalize fairly well towards new data. The speech utterances are coded using the AMR-NB and AMR-WB codecs. The coded speech data is processed in 30-ms frames at 15-ms intervals in the dysarthric speech classification system. Using openSMILE, two sets of acoustic features (openSMILE-1 and openSMILE-2) are extracted from every coded speech utterance of the TORGO and UA-Speech databases. From every coded speech utterance, voiced segments are extracted using the SEDREAMS method (Drugman et al., 2012b). From every voiced frame of coded speech utterance, glottal flow waveforms are estimated using QCP and DNN-GIF. From the glottal flow waveforms of every utterance, two sets of glottal parameters (Glottal-1 and Glottal-2) are extracted. The time- and frequency-domain glottal parameters extracted using the QCP and DNN-GIF methods are denoted as 'Glottal-1 (QCP)' and 'Glottal-1 (DNN-GIF)' respectively. The PCA-based glottal parameters extracted using the QCP and DNN-GIF are denoted as 'Glottal-2 (QCP)' and 'Glottal-2 (DNN-GIF)' respectively. Both acoustic and glottal features are individually normalized by subtracting the global mean and dividing by the global standard deviation. The sizes of each of the feature sets are reduced by the SFFS algorithm on the validation data. The process of feature extraction and reduction is carried out separately for both NB- and WB-coded speech data. Separate sets of SVM classifiers are developed for both NB- and WB-coded speech using acoustic and glottal feature sets both individually and combined. Also, SVM classifiers are developed for both the non-reduced and reduced feature sets. The SVM classifiers are trained using the Gaussian, radial basis function kernel. The optimal values of kernel parameter γ and penalty parameter C are chosen based on grid search with C and γ, varying from 10^{-3} to 10^{3} in multiples of 10. The pair (C, γ) is selected, which resulted in the highest classification accuracy on the validation data.

A leave-one-subject-out (LOSO) cross-validation strategy is used to determine the classification accuracy on the training data. In this strategy, one speaker is used at every fold for validation and all other speakers are used for training. The cross-validation process is then repeated with each speaker used exactly once as the validation data. The classification accuracies obtained at all folds are averaged to obtain the final accuracy. The accuracy is calculated by comparing for the speech utterance, the predicted label (*dysarthria* vs. *healthy*) to the known condition of the individual. The classification accuracy (also called the *unweighted average recall*) is computed as the ratio of number of correctly classified speech utterances to the total number of speech utterances.

In addition to the classification accuracy, the performance of classifiers is assessed using area under receiver operating characteristics (ROC) curves, termed as AUC. The ROC is a graph showing the performance of a classifier at various classification thresholds. This graph plots the true positive rate against the false positive rate at different classification thresholds. AUC measures the area under the ROC curve. AUC represents the degree of separability of a classifier. The value of AUC ranges from 0 to 1. If AUC is 0, then the classifier predictions are 100% wrong and if AUC is 1, then the classifier predictions are 100% right.

Table 3
Classification accuracies obtained for both NB- and WB-coded speech using the non-reduced and reduced feature sets of the TORGO database.

Feature set (NB-coded)	Classification accuracy	
	Without feature selection (%)	With feature selection (%)
OpenSMILE-1	63.87	77.71
OpenSMILE-2	64.49	82.79
Glottal-1 (QCP)	43.52	64.12
Glottal-2 (QCP)	60.48	63.60
Glottal-1 (DNN-GIF)	54.01	72.76
Glottal-2 (DNN-GIF)	63.75	77.34
OpenSMILE-1 + Glottal-1 (QCP)	56.66	79.50
OpenSMILE-2 + Glottal-1 (QCP)	62.49	83.59
OpenSMILE-1 + Glottal-2 (QCP)	65.52	79.19
OpenSMILE-2 + Glottal-2 (QCP)	64.67	82.93
OpenSMILE-1 + Glottal-1 (DNN-GIF)	61.70	81.71
OpenSMILE-2 + Glottal-1 (DNN-GIF)	63.47	84.36
OpenSMILE-1 + Glottal-2 (DNN-GIF)	67.49	81.62
OpenSMILE-2 + Glottal-2 (DNN-GIF)	64.03	84.82
Feature set (WB-coded)	Without feature selection (%)	With feature selection (%)
OpenSMILE-1	67.16	83.54
OpenSMILE-2	67.27	88.61
Glottal-1 (QCP)	50.58	66.19
Glottal-2 (QCP)	63.09	68.13
Glottal-1 (DNN-GIF)	59.54	70.57
Glottal-2 (DNN-GIF)	56.47	71.24
OpenSMILE-1 + Glottal-1 (QCP)	59.35	83.88
OpenSMILE-2 + Glottal-1 (QCP)	66.02	88.68
OpenSMILE-1 + Glottal-2 (QCP)	67.34	87.23
OpenSMILE-2 + Glottal-2 (QCP)	67.52	89.62
OpenSMILE-1 + Glottal-1 (DNN-GIF)	68.35	85.58
OpenSMILE-2 + Glottal-1 (DNN-GIF)	67.56	89.81
OpenSMILE-1 + Glottal-2 (DNN-GIF)	63.66	85.17
OpenSMILE-2 + Glottal-2 (DNN-GIF)	66.59	89.91

Table 4
Classification accuracies obtained for both NB- and WB-coded speech using the non-reduced and reduced feature sets of the UA-Speech database.

Feature set (NB-coded)	Classification accuracy	
	Without feature selection (%)	With feature selection (%)
OpenSMILE-1	90.42	91.18
OpenSMILE-2	95.11	95.25
Glottal-1 (QCP)	69.75	74.31
Glottal-2 (QCP)	67.60	68.58
Glottal-1 (DNN-GIF)	78.06	78.51
Glottal-2 (DNN-GIF)	81.49	80.54
OpenSMILE-1 + Glottal-1 (QCP)	87.66	91.70
OpenSMILE-2 + Glottal-1 (QCP)	94.40	95.64
OpenSMILE-1 + Glottal-2 (QCP)	88.39	91.22
OpenSMILE-2 + Glottal-2 (QCP)	94.85	95.58
OpenSMILE-1 + Glottal-1 (DNN-GIF)	89.63	91.82
OpenSMILE-2 + Glottal-1 (DNN-GIF)	95.17	95.81
OpenSMILE-1 + Glottal-2 (DNN-GIF)	88.03	91.99
OpenSMILE-2 + Glottal-2 (DNN-GIF)	95.20	96.07
Feature set (WB-coded)	Without feature selection (%)	With feature selection (%)
OpenSMILE-1	85.91	88.03
OpenSMILE-2	95.45	95.57
Glottal-1 (QCP)	69.17	77.21
Glottal-2 (QCP)	73.43	75.64
Glottal-1 (DNN-GIF)	76.71	78.13
Glottal-2 (DNN-GIF)	78.46	78.79
OpenSMILE-1 + Glottal-1 (QCP)	85.83	88.23
OpenSMILE-2 + Glottal-1 (QCP)	95.02	95.87
OpenSMILE-1 + Glottal-2 (QCP)	85.45	88.76
OpenSMILE-2 + Glottal-2 (QCP)	95.37	96.07
OpenSMILE-1 + Glottal-1 (DNN-GIF)	85.08	89.67
OpenSMILE-2 + Glottal-1 (DNN-GIF)	95.60	96.16
OpenSMILE-1 + Glottal-2 (DNN-GIF)	86.00	89.50
OpenSMILE-2 + Glottal-2 (DNN-GIF)	95.40	96.38

In order to analyze the effectiveness of the classifiers developed with the baseline features (openSMILE-1 and openSMILE-2) and different combinations of the baseline (openSMILE-1 and openSMILE-2) and glottal features (Glottal-1 and Glottal-2), statistical tests were conducted using Cochran's Q test (Daniel, 1978). Cochran's Q test can be regarded as a generalized version of McNemar's test that can be applied to compare the performance of multiple classifiers. Cochran's Q tests the null hypothesis that there is no difference between the classification accuracies. Cochran's Q test was performed separately on all the classifiers of NB- and WB-coded speech.

3.4. Results

Tables 3 and 4 show the average classification accuracies of the leave-one-subject-out cross-validation for both NB- and WB-coded speech using the reduced and non-reduced feature sets of the TORGO and UA-Speech databases. In comparing the classification accuracies for all types of feature sets of NB- and WB-coded speech, it can be observed that the usage of the reduced feature sets results in better accuracy compared to the non-reduced feature sets for both TORGO and UA-Speech. From the table, it can be observed that with more than 80% classification accuracy (except for openSMILE-1 of NB-coded speech with 77.71% of TORGO), the two sets of openSMILE-based features have better classification accuracy than the glottal parameters after feature selection for both NB- and WB-coded speech. The classification accuracies of the two sets of glottal parameters (Glottal-1 and Glottal-2) obtained from QCP and DNN-GIF vary between 63 - 80% after feature selection for both NB- and WB-coded speech of TORGO and UA-Speech. This indicates that the glottal parameters contain discriminative information, important for the classification of dysarthric speech. In comparing the classification accuracies of the two sets of glottal parameters obtained from QCP and DNN-GIF for both NB- and WB-coded speech, it can be observed that the glottal parameters obtained from DNN-GIF results in improved accuracy. This indicates that DNN-GIF can accurately estimate the glottal flow under the coded condition and, hence, the resulting glottal parameters (both Glottal-1 and Glottal-2) obtained from DNN-GIF are more effective in the identification of dysarthric speech. Among the two sets of glottal parameters, Glottal-2 (i.e., the PCA-based glottal parameters) results in better accuracy compared to Glottal-1 (i.e., time- and frequency-domain glottal parameters) in most of the cases.

By combining the glottal features with the openSMILE features, the classification accuracies improve after feature selection for both NB- and WB-coded speech of TORGO and UA-Speech. This shows that the glottal fetaures contain complementary information, which results in the improvement of accuracies when combined with the openSMILE features. On closer analysis of the results in Tables 3 and 4, it can be observed that the classification accuracies obtained from the combination of the openSMILE features and the DNN-GIF-computed glottal features are higher than the combination of the openSMILE features and the QCP-computed glottal parameters for both NB- and WB-coded speech (except OpenSMILE-1 + Glottal-2 (QCP) for WB-coded speech of the TORGO database).

Table 5 shows the AUC values obtained from the reduced feature sets of TORGO and UA-Speech for both NB- and WB-coded speech. It can be observed from the table that the AUC values exhibit similar patterns to the classification accuracies for both TORGO and UA-Speech. From Tables 3 and 4, it can be observed that the classifier developed with the OpenSMILE-2 + Glottal-2 (DNN-GIF) feature set has the highest classification accuracy for both NB- and WB-coded speech. From the Cochran's Q test performed separately on all the classifiers of NB- and WB-coded speech, it is observed that the null hypothesis was rejected at $\alpha = 0.0056$, *degrees of freedom* = 9. This indicates that all classifiers

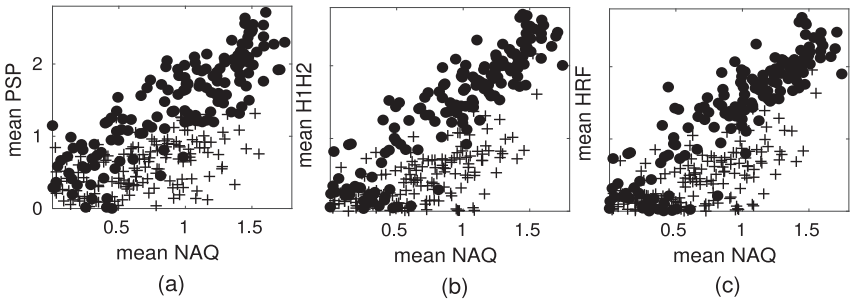

Fig. 4. Scatter plots between pairs of glottal features for dysarthric and normal speech of randomly selected 100 utterances from the TORGO database. (a) mean NAQ and mean PSP, (b) mean NAQ and mean H1H2, and (c) mean NAQ and mean HRF. Round mark indicates normal speech and '+' mark indicates dysarthric speech.

Table 5
Area under ROC performances of different classifiers obtained from the reduced feature sets of the TORGO and UA-Speech databases for both NB- and WB-coded speech.

Feature set (NB-coded)	Area under ROC (AUC)	
	TORGO	UA-Speech
OpenSMILE-1	0.8081	0.9601
OpenSMILE-2	0.8705	0.9812
Glottal-1 (QCP)	0.7046	0.8098
Glottal-2 (QCP)	0.7197	0.7349
Glottal-1 (DNN-GIF)	0.8041	0.8633
Glottal-2 (DNN-GIF)	0.7632	0.8822
OpenSMILE-1 + Glottal-1 (QCP)	0.8663	0.9624
OpenSMILE-2 + Glottal-1 (QCP)	0.8981	0.9831
OpenSMILE-1 + Glottal-2 (QCP)	0.8290	0.9642
OpenSMILE-2 + Glottal-2 (QCP)	0.9026	0.9840
OpenSMILE-1 + Glottal-1 (DNN-GIF)	0.8690	0.9688
OpenSMILE-2 + Glottal-1 (DNN-GIF)	0.8902	0.9881
OpenSMILE-1 + Glottal-2 (DNN-GIF)	0.8615	0.9654
OpenSMILE-2 + Glottal-2 (DNN-GIF)	0.8978	0.9893
Feature set (WB-coded)	TORGO	UA-Speech
OpenSMILE-1	0.8955	0.9510
OpenSMILE-2	0.9445	0.9742
Glottal-1 (QCP)	0.6761	0.8311
Glottal-2 (QCP)	0.7702	0.8165
Glottal-1 (DNN-GIF)	0.7183	0.8640
Glottal-2 (DNN-GIF)	0.7603	0.8388
OpenSMILE-1 + Glottal-1 (QCP)	0.8947	0.9553
OpenSMILE-2 + Glottal-1 (QCP)	0.9499	0.9776
OpenSMILE-1 + Glottal-2 (QCP)	0.8975	0.9574
OpenSMILE-2 + Glottal-2 (QCP)	0.9460	0.9757
OpenSMILE-1 + Glottal-1 (DNN-GIF)	0.9014	0.9536
OpenSMILE-2 + Glottal-1 (DNN-GIF)	0.9515	0.9835
OpenSMILE-1 + Glottal-2 (DNN-GIF)	0.9086	0.9587
OpenSMILE-2 + Glottal-2 (DNN-GIF)	0.9546	0.9843

are not effectively equal in identifying the presence of dysarthria. Later, a pairwise Cochran's Q test was conducted between the best performing classifier developed with the OpenSMILE-2 + Glottal-2 (DNN-GIF) feature set and all other classifiers. The test showed that the best performing classifier was significantly different to all other classifiers at $p < 0.005$, except for the classifier developed with the OpenSMILE-2 + Glottal-1 (DNN-GIF) feature set for both NB- and WB-coded speech.

Fig. 4 shows the scatter plots of pairs of glottal features for dysarthric and normal speech of randomly selected 100 utterances from the TORGO database. The glottal features are computed from WB-coded speech using DNN-GIF. Even though there is slight overlap in scatter plots of pairs of glottal features, dysarthric speech can be separated from normal speech with appropriate decision boundaries.

4. Conclusion

A new dysarthric speech classification method from coded telephone speech is proposed using glottal features. The proposed method utilizes SVM to predict *dysarthric/healthy* labels by using the acoustic and glottal features extracted from coded speech. Two sets of acoustic features are extracted using the openSMILE toolkit and two sets of glottal features are extracted from glottal flow waveform. The glottal flow waveform is obtained from coded telephone speech (coded with two standardized speech codecs - AMR-NB and AMR-WB) using QCP and the recently proposed DNN-GIF method. Experimental results show that the glottal parameters resulted in fairly good classification accuracy (63–77%) for both NB- and WB-coded speech. From the two glottal parameter sets, the PCA-based glottal parameters resulted in better accuracy than the conventional time- and frequency-domain parameters. The results showed that the glottal parameters obtained from DNN-GIF lead to better accuracy compared to the parameters obtained from QCP. This proves the effectiveness of DNN-GIF in the estimation of the glottal flow from coded telephone speech. Experiments also showed that combining the glottal parameters with the openSMILE features results in improved classification accuracies. This improvement in classification accuracy was shown to be statistically significant.

To the best of our knowledge, the current study is the first detailed investigation on dysarthric speech classification using glottal features from coded telephone speech. The proposed method showed the effectiveness of glottal parameters obtained from the recently proposed DNN-GIF method in dysarthric speech classification. The proposed method also showed consistent performance for different combinations of the openSMILE and glottal features, and for both NB- and WB-coded speech. Possible future works are as follows. Apart from the AMR-NB and AMR-WB codecs, the proposed method can be evaluated using recent codecs, for example, Enhanced Voice Services (EVS) codec (3GPP TS 26.445, 2014). The proposed method can be extended for the speech-based telemonitoring of different neuro-motor disorders such as Parkinson's disease, Alzheimer's disease, and ALS. Apart from neuro-motor disorders, the proposed method can be utilized for different paralinguistic tasks such as the recognition of emotion, and speaker states and traits under the coded condition.

Acknowledgment

This research has been funded by the Academy of Finland (project no. 312490).

Declaration of interests

No conflict.

References

3GPP TS 26.090, version 10.1.0, 2011. Adaptive Multi-rate (AMR) Speech Codec, Transcoding Functions. Technical Report. 3rd Generation Partnership Project.
3GPP TS 26.445, 2014. EVS Codec Detailed Algorithmic Description; 3GPP Technical Specification (Release 12).
Airaksinen, M., Juvela, L., Bollepalli, B., Yamagishi, J., Alku, P., 2018. A comparison between STRAIGHT, glottal, and sinusoidal vocoding in statistical parametric speech synthesis. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (9), 1658–1670.
Airaksinen, M., Raitio, T., Alku, P., 2015. Noise robust estimation of the voice source using a deep neural network. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5137–5141.
Airaksinen, M., Raitio, T., Story, B., Alku, P., 2014. Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (3), 596–607.
Airas, M., Pulakka, H., Bäckström, T., Alku, P., 2005. A toolkit for voice inverse filtering and parametrisation. In: Proc. Interspeech, pp. 2145–2148.

Alku, P., 1992. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Commun. 11 (2–3), 109–118.
Alku, P., 2011. Glottal inverse filtering analysis of human voice production a review of estimation and parameterization methods of the glottal excitation and their applications. Sadhana 623–650.
Alku, P., Bäckström, T., Vilkman, E., 2002. Normalized amplitude quotient for parameterization of the glottal flow. J. Acoust. Soc. Am. 112 (2), 701–710.
Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A.-M., Story, B.H., 2013. Formant frequency estimation of high-pitched vowels using weighted linear prediction. J. Acoust. Soc. Am. 134 (2), 1295–1313.
Arias-Vergara, T., Vásquez-Correa, J., Orozco-Arroyave, J., Nöth, E., 2018. Speaker models for monitoring Parkinson's disease progression considering different communication channels and acoustic conditions. Speech Commun. 101, 11–25.
Childers, D., Naik, J., Larar, J., Krishnamurthy, A., Moore, G.R., 1983. Vocal Fold Physiology, Biomechanics, Acoustics and Phonatory Control. The Denver Center for The Performing Arts, Denver, pp. 202–220.
Childers, D.G., Lee, C.K., 1991. Vocal quality factors: analysis, synthesis, and perception. J. Acoust. Soc. Am. 90 (5), 2394–2410.
Cinnéide, A.O., Dorran, D., Gainza, M., Coyle, E., 2010. Exploiting glottal formant parameters for glottal inverse filtering and parameterization. In: Proc. Interspeech.
Constantinescu, G., Theodoros, D., Russell, T., Ward, E., Wilson, S., Wootton, R., 2010. Assessing disordered speech and voice in Parkinson's disease: a telerehabilitation application. Int. J. Lang. Commun.Disord. 45 (6), 630–644.
Daniel, W.W., 1978. Applied Nonparametric Statistics. Houghton Mifflin.
De Bodt, M.S., Hernández-Díaz Huici, M.E., Van De Heyning, P.H., 2002. Intelligibility as a linear combination of dimensions in dysarthric speech. J. Commun. Disord. 35 (3), 283–292.
Deller, J.R., Liu, M.S., Ferrier, L.J., Robichaud, P., 1993. The whitaker database of dysarthric (cerebral palsy) speech. J. Acoust. Soc. Am. 93 (6), 3516–3518.
Dibazar, A.A., Narayanan, S., Berger, T.W., 2002. Feature analysis for automatic detection of pathological speech. In: Proc. Joint EMBS/BMES Conference. 182–182.
Doyle, P.C., Leeper, H.A., Kotler, A.L., Thomas-Stonell, N., O'Neill, C., Dylke, M.C., Rolls, K., 1997. Dysarthric speech: a comparison of computerized speech recognition and listener intelligibility. J. Rehabil. Res. Dev. 34 (3), 309–316.
Drugman, T., Bozkurt, B., Dutoit, T., 2009. Complex cepstrum-based decomposition of speech for glottal source estimation. In: Proc. interspeech, pp. 116–119.
Drugman, T., Bozkurt, B., Dutoit, T., 2012. A comparative study of glottal source estimation techniques. Comput. Speech Lang. 25 (1), 20–34.
Drugman, T., Dutoit, T., 2012. The deterministic plus stochastic model of the residual signal and its applications. IEEE Trans. Audio Speech Lang. Process. 20 (3), 968–981.
Drugman, T., Thomas, M., Gudnason, J., Naylor, P., Dutoit, T., 2012. Detection of glottal closure instants from speech signals: a quantitative review. IEEE Trans. Audio Speech Lang. Process. 20 (3), 994–1006.
Duffy, J.R., 2012. Motor Speech Disorders: Substrates, Differential Diagnosis, and Management, Third ed. Elsevier Health Sciences.
Enderby, P.M., 1983. Frenchay Dysarthria Assessment. College Hill Press, San Diego.
Eyben, F., Weninger, F., Gross, F., Schuller, B., 2013. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: Proc. ACM International Conference on Multimedia, pp. 835–838.
Falk, T.H., Chan, W.-Y., Shein, F., 2012. Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. Speech Commun. 54, 622–631.
Gallardo, L.F., Wagner, M., Möller, S., 2014. I-vector speaker verification based on phonetic information under transmission channel effects. In: Interspeech, pp. 696–700.
Gillespie, S., Logan, Y.-Y., Moore, E., Laures-Gore, J., Russell, S., Patel, R., 2017. Cross–database models for the classification of dysarthria presence. In: Proc. Interspeech, pp. 3127–3131.
Järvinen, K., 2000. Standardisation of the adaptive multi-rate codec. In: Proc. European Signal Processing Conference (EUSIPCO).
Kent, R.D., 1992. Intelligibility in Speech Disorders: Theory, Measurement, and Management, Vol. 1. John Benjamins Publishing.
Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T., Watkin, K., Frame, S., 2008. Dysarthric speech database for universal access research. In: Proc. Interspeech, pp. 1741–1744.
Kim, J., Kumar, N., Tsiartas, A., Li, M., Narayanan, S.S., 2015. Automatic intelligibility classification of sentence-level pathological speech. Comput. Speech. Lang. 29, 132–144.
Klumpp, P., Janu, T., Arias-Vergara, T., Correa, J.C.V., Orozco-Arroyave, J.R., Nöth, E., 2017. Apkinson - a mobile monitoring solution for Parkinson's disease. In: Proc. Interspeech, pp. 1839–1843.
Lei, H., Lopez-Gonzalo, E., 2009. Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition. In: Proc. Interspeech, pp. 2323–2326.
Little, M.A., McSharry, P.E., Hunter, E.J., Spielman, J., Ramig, L.O., 2009. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Trans. Biomed. Eng. 56 (4), 1015–1022.
Mandal, I., Sairam, N., 2013. Accurate telemonitoring of Parkinson's disease diagnosis using robust inference system. Int. J. Med. Inform. 82 (5), 359–377.
Medennikov, I., Prudnikov, A., Zatvornitskiy, A., 2016. Improving English conversational telephone speech recognition. In: Proc. Interspeech.
Menendez-Pidal, X., Polikoff, J.B., Peters, S.M., Leonzjo, J.E., Bunnell, H., 1996. The Nemours database of dysarthric speech. In: Proc. International Conference on Spoken Language Processing (ICSLP), pp. 1962–1965.
Narendra, N.P., Airaksinen, M., Alku, P., 2017. Glottal source estimation from coded telephone speech using a deep neural network. In: Proc. Interspeech, pp. 3931–3935.
Narendra, N.P., Airaksinen, M., Story, B., Alku, P., 2018. Estimation of the glottal source from coded telephone speech using deep neural networks. Speech Commun. doi:10.1016/j.specom.2018.12.002.
Narendra, N.P., Alku, P., 2018. Dysarthric speech classification using glottal features computed from non-words, words and sentences. In: Proc. Interspeech. 3403–3307.
Orozco-Arroyave, J.R., Hönig, F., Arias-Londono, J.D., Vargas-Bonilla, J.F., Skodda, S., Rusz, J., Nöth, E., 2014. Automatic detection of Parkinson's disease from words uttered in three different languages. In: Proc. Interspeech, pp. 1573–1577.
Raitio, T., Suni, A., Vainio, M., Alku, P., 2013. Comparing glottal flow-excited statistical parametric speech synthesis methods. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7830–7834.
Ramezani, H., Khaki, H., Erzin, E., Akan, O.B., 2017. Speech features for telemonitoring of Parkinson's disease symptoms. In: Proc. International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3801–3805.
Reunanen, J., 2003. Overfitting in making comparisons between variable selection methods. J. Mach. Learn. Res. 3, 1371–1382.
Rudzicz, F., 2009. Phonological features in discriminative classification of dysarthric speech. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4605–4608.
Rudzicz, F., Namasivayam, A.K., Wolff, T., 2012. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Lang. Resour. Eval. 46 (4), 523–541.
Sakar, B.E., Serbes, G., Sakar, C.O., 2017. Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease. PLoS ONE 12 (8), 1–18.
Schuller, B., Steidl, S., Batliner, A., 2009. The INTERSPEECH 2009 emotion challenge. In: Proc. Interspeech, pp. 312–315.
Thomas, M., Gudnason, J., Naylor, P., 2009. Data-driven voice source waveform modelling. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3965–3968.
Tsanas, A., Little, M.A., McSharry, P.E., Ramig, L.O., 2010. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. IEEE Trans. Biomed. Eng. 57 (4), 884–893.
Van Nuffelen, G., Middag, C., De Bodt, M., Martens, J.-P., 2009. Speech technology-based assessment of phoneme intelligibility in dysarthria. Int. J. Lang. Commun.Disord. 44 (5), 716–730.
Vásquez-Correa, J., Arias-Vergara, T., Orozco-Arroyave, J.R., Vargas-Bonilla, J.F., no, J.D.A.-L., Nöth, E., 2015. Automatic detection of Parkinson's disease from continuous speech recorded in non-controlled noise conditions. In: Proc. Interspeech, pp. 105–109.
Wong, D., Markel, J., Gray Jr, A., 1979. Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Trans. Audio Speech Lang. Process. 27 (4), 350–355.
Wrench, A., 1999. The MOCHA-TIMIT articulatory database.
Yorkston, K.M., Beukelman, D.R., 1981. Assessment of Intelligibility of Dysarthric Speech. Tigard, OR: C.C. Publications.