Computer Speech & Language 65 (2021) 101117

Contents lists available at ScienceDirect

Computer Speech & Language

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

Automatic assessment of intelligibility in speakers with dysarthria 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

Article History:
Received 10 September 2019
Revised 2 April 2020
Accepted 19 May 2020
Available online 2 June 2020

Keywords:
Dysarthric speech
Glottal features
Glottal inverse filtering
Glottal source estimation
Opensmile
Support vector machine
Coded telephone speech

ABSTRACT

In clinical practice, assessment of intelligibility in speakers with dysarthria is performed by speech-language pathologists through auditory perceptual tests which demand patients' presence at hospital and involve time-consuming examinations. Frequent clinical monitoring can be costly and logistically inconvenient both for patients and medical experts. Here, we aim to automate the procedure of assessment of intelligibility in dysarthric speakers with an objective, speech-based method that can be employed in a telescreening application. The proposed method predicts the level of intelligibility in dysarthric speakers using four levels of speech intelligibility (very low, low, mediocre and high). The study compares several automatic methods to assess the intelligibility level in speakers with dysarthria by utilizing information generated at the level of the vocal folds through glottal features and by using coded telephone speech (i.e. speech that is used in telescreening applications). In addition to the glottal features, the openSMILE features are used as acoustic baseline features. Using features obtained from coded speech utterances and the corresponding intelligibility level labels, multiclass-support vector machine (SVM) classifiers are trained. A separate set of multiclass-SVMs are trained using both individual glottal and acoustic features as well as their combinations. Coded telephone speech is generated with the adaptive multi-rate codec with two operational bandwidths (narrowband and wideband), from utterances of an open database of dysarthric speech (Universal Access-Speech). Experimental results showed good classification accuracies for the glottal features, indicating their effectiveness in the intelligibility level assessment in speakers with dysarthria even in the challenging coded condition. Improvement in classification accuracy was obtained when the glottal features were combined with the openSMILE acoustic features, which validate the complimentary nature of the glottal features.

1. Introduction

Dysarthria is a neuro-motor disorder resulting in neurological damage to muscular control of the speech production mechanism (Doyle et al., 1997). Dysarthria is generally a result of amyotropic lateral sclerosis (ALS), Parkinson's disease (PD), brain tumor, brain injury, or cerebral palsy. Dysarthric speech is often characterized by imprecise articulation, variable speech rate, and irregular speech prosody - factors that lead to reduction in speech intelligibility (Duffy, 2012). The intelligibility assessment of dysarthric speech can be considered as a diagnostic step, which is crucial to understand the patient's progression. This helps to

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

https://doi.org/10.1016/j.csl.2020.101117
0885-2308/

consider critical decisions related to the course of medical treatment. This study focuses on the automatic assessment of the *level of speech intelligibility* (very low/low/mediocre/high) in speakers with dysarthria.

Depending on the severity of dysarthria, the intelligibility of speech can vary from near-normal to unintelligible (Polikoff and Bunnel, 1999). The assessment of intelligibility of speech can be carried out using a conventional method, which includes subjective intelligibility tests performed by speech-language pathologists (Kent et al., 1989). The subjective speech assessment tests are, however, expensive, time consuming and suffer from non-reproducibility and subjectivity (De Bodt et al., 2002; Van Nuffelen et al., 2009). By contrast, speech-based objective assessment benefits from its low cost and being replicative, consistent and without any subjective bias (Carmichael, 2007). These factors motivate the design of an objective assessment method to estimate the intelligibility level in dysarthric speakers by using their speech signals.

Objective assessment of speech intelligibility in speakers with dysarthria can be reliable and economical, and it can be utilized to carry out the diagnosis on a regular basis by easily collecting speech samples from patients (Constantinescu et al., 2010). Since collecting speech samples is noninvasive, the speech-based evaluation can be carried out without any conventional medical equipment and it can be readily integrated into a telescreening application (Tsanas et al., 2010; Sakar et al., 2017). Speech-based telescreening can supplement conventional medical diagnosis methods and reduce the burden of frequent visits to hospital (Goetz et al., 2009). In literature, the speech-based assessment of neuro-motor disorders for telescreening and telemonitoring applications has been studied mainly related to Parkinson's disease (Tsanas et al., 2010; Sakar et al., 2017; Little et al., 2009; Mandal and Sairam, 2013; Klumpp et al., 2017). In Mandal and Sairam (2013); Klumpp et al. (2017), telemonitoring frameworks were proposed for the regular examination of Parkinson's disease. Tsanas et al. Tsanas et al. (2010) explored clinically useful features such as detrended fluctuation analysis (DFA), recurrence period density entropy (RPDE) and pitch period entropy (PPE) for the remote monitoring of the progression of Parkinson's disease. Most of the existing works on speech-based telescreening primarily utilize clean speech, which is free from coding and other degrading effects. Therefore, in order to build an efficient telescreening system, there is a need to perform the assessment of intelligibility in speakers with dysarthria in more challenging situations. In the present study, an intelligibility assessment system that works with coded dysarthric speech - data that is most appropriate for telescreening applications - is studied.

Even though coded speech is used in this article, a real remote telescreening system to screen dysarthric speakers is not directly addressed in the current study. In addition to speech compression, there are namely also other issues (such as channel degradation and environment noise) which need to be studied for the method to be implementable in a real telescreening application. However, the current study can be considered as a preliminary step towards developing such a telescreening system. In principle, the proposed method can be used by clinicians as the first level of screening of dysarthric speakers, as it predicts the intelligibility level in speakers with dysarthria.

Previous studies in objective assessment of dysarthric speech mainly focus on the binary classification task (i.e., identify a given speech signal as dysarthric or healthy) (Rudzicz, 2009; Kim et al., 2015). Previous methods on the assessment of dysarthric speech intelligibility have mainly explored spectral features (e.g. formants, line spectral frequencies (LSFs), Mel-frequency cepstral coefficients (MFCCs)), prosody features (e.g. phone duration, energy, F0, pitch contour), and voice quality features (e.g. jitter, shimmer, harmonic-to-noise ratio) (Laaridh et al., 2017; Kim and Kim, 2012; Falk et al., 2012; Martínez et al., 2013; Kadi et al., 2016). However, features describing the glottal flow (i.e., the source of voiced speech generated by the vocal folds) have not been used previously in the intelligibility assessment of dysarthric speech. These so-called *glottal features* have, however, been used recently in dysarthric speech classification studies (Gillespie et al., 2017; Narendra and Alku, 2018; 2019). Gillespie et al. (2017) explored glottal parameters in combination with prosodic and spectral features for the classification of dysarthric speech. Recent studies in the classification of dysarthria by the present authors Narendra and Alku (2018, 2019) illustrated the importance of the features describing the glottal source, as well as indicated that these features may provide complementary information to the existing acoustic features.

Telescreening of dysarthric speech can be performed at the transmitter-end, such as a smartphone, where speech is processed directly (Klumpp et al., 2017) or at the receiver-end where the transmitted speech is remotely processed by a dedicated diagnosis system (Mandal and Sairam, 2013). In both of these scenarios, assessment of dysarthric speech using glottal features is highly challenging, as glottal inverse filtering (GIF) methods (Alku, 2011) that are required to estimate the glottal flow from speech are observed to be highly vulnerable even to slight degradation in the speech utterance (Airaksinen et al., 2015; Narendra et al., 2017). In the first scenario, speech is typically recorded by the smartphone's built-in microphone, whose frequency response is poor compared to high-quality condenser microphones (e.g. the phase response is non-linear). This results in phase distortion of the recorded speech signal, which in turn is known to severely degrade the estimation of the glottal flow by GIF (Wong et al., 1979). In the second scenario, speech may be distorted even more by additional factors such as transmission errors, band-pass filtering, and low bit-rate speech coding. In this work, assessment of intelligibility level in speakers with dysarthria is performed under one of the degradation factors, i.e., speech coding.

In order to improve the robustness of the GIF-based estimation of glottal flows under coded condition, a deep neural net-based glottal inverse filtering method, DNN-GIF, was recently proposed (Narendra et al., 2019). In DNN-GIF, a deep neural net (DNN) is used to estimate glottal flows using spectral features computed from coded telephone speech. In Narendra et al. (2019), DNN-GIF was shown to be the best performing technique in comparison to state-of-the-art GIF techniques under coded condition. The present study explores the glottal features obtained using DNN-GIF under coded condition for the assessment of intelligibility level in speakers with dysarthria. In addition, the study aims to investigate whether the glottal features provide complementary information to the existing acoustic features which can be helpful in the intelligibility assessment. The

assessment of intelligibility in speakers with dysarthria is carried out from coded telephone speech with two operational bandwidths, narrowband (300 Hz − 3.4 kHz) and wideband (50 Hz − 7 kHz).

This article proposes a new method to estimate the intelligibility level in speakers with dysarthria from coded telephone speech (i.e. from speech used in telescreening applications). The study specifically analyzes the importance of glottal features that are extracted from coded telephone speech for the intelligibility level assessment. Glottal source signals are estimated using two methods: (1) the recently proposed deep neural net-based glottal inverse filtering method (Narendra et al., 2019) (which was shown to be the most accurate GIF under coded condition) and the quasi-closed phase analysis (QCP) method (Airaksinen et al., 2014) (which was shown to be the most accurate GIF method for clean speech). Glottal source signals estimated by both of these methods are parameterized into two sets: (1) time- and frequency-domain features and (2) principal component analysis (PCA)-based features. In addition to the glottal features, two sets of baseline acoustic features extracted using the openSMILE toolkit (Eyben et al., 2013) are used. The acoustic features extracted using the openSMILE toolkit have been used as the baseline features in different paralinguistic tasks such as in emotion recognition and in detection of speaker traits and states (e.g. depression, stress, styling, confidence, deception) (Schuller et al., 2009; 2012; 2015). Multiclass support vector machine (SVM) classifiers are trained using features obtained from coded telephone speech as input and the corresponding intelligibility level label as output. The study compares the glottal features (both alone and combined with the baseline openSMILE features) in assessing the level of intelligibility in speakers with dysarthria using a freely available database (Kim et al., 2008).

This paper is organized as follows. In Section II, the importance of the glottal source in the assessment of speech intelligibility in speakers with dysarthria is discussed. The methods used in glottal inverse filtering and in extraction of glottal features are described in Section III. Section IV describes the proposed intelligibility level assessment system. The details of the dysarthric speech database and the results of the experiments are given in Section V. Conclusions of the proposed method and possible extensions to the present study are summarized in Section VI.

2. Glottal source in dysarthric speech

Glottal source is the pulsatile air flow waveform produced by the oscillation of the vocal folds due to interaction of the mechanical tissue properties with respiratory pressure (Rabiner and Schafer, 1978). In production of speech, the sequence of resonators formed by the vocal tract cavities filters the input glottal source and the resulting flow passing through the lips is transformed into the speech pressure waveform. It is practically not possible to measure directly and non-invasively the glottal flow at the level of the vocal folds. However, the glottal source can be estimated indirectly from the microphone speech signal using computational inversion known as glottal inverse filtering (Alku, 2011). Glottal source information, estimated by GIF, has been shown to be important in different speech research areas such as in analysis of voice quality (Campbell and Mokhtari, 2003), emotion (Airas and Alku, 2006), prosody (Airas et al., 2007) and singing voices (Arroabarren and Carlosena, 2006), as well as in speaker identification (Kinnunen and Alku, 2009) and in speech synthesis (Raitio et al., 2011). Moreover, in the context of pathological voice, glottal source has been used for classification of disordered speech (Chien et al., 2017), Parkinson's disease (Hanratty et al., 2016), and Alzheimer's disease (Fraser et al., 2016). However, in the context of assessment of intelligibility level in speakers with dysarthria, it is worth noting that the glottal source has not been taken advantage of.

In dysarthria, the normal function of motor component of speech production is affected, and as a result, deviation in the nature of vibration of the vocal folds can be observed in dysarthric speech compared to healthy speech (Duffy, 2000). The vibratory nature of the vocal folds is affected at different levels depending on the severity of dysarthria. Dysarthria affects vocal fold vibration mainly in two respects: by changing the rate of vocal fold vibration and by altering the shape of the glottal airflow pulse generated by the vocal folds. Pitch and jitter (variability of F0 across several cycles of vibration) have been explored for the assessment of dysarthria (Kim and Kim, 2012; Falk et al., 2012; Kadi et al., 2014). In addition, perceptual evaluation studies of neuro-motor diseases (including dysarthric speech due to cerebral palsy and Parkinson's disease) have indicated the deterioration in voice quality factors (such as hoarseness, breathiness) with progression of the disease severity (Skodda et al., 2013; Lansford et al., 2014). Moreover, previous studies analysing healthy speech have demonstrated that changes in the glottal pulse shape affect voice quality factors both in male and female speakers (Alku and Vilkman, 1996; Alku et al., 2006) and parameters representing the pulse shape can be used to differentiate voice qualities. Based on these studies, we can hypothesize that the shape of the glottal flow pulse may contain important information for automatic assessment of intelligibility level in speakers with dysarthria.

The above hypothesis is demonstrated in Fig. 1 by showing examples of speech signals and the corresponding glottal flows estimated from parts of the vowel [a] in three speech intelligibility levels in dysarthria: high (a), mediocre (b), and low (c). The speech signals are clean (i.e., not in coded condition) and the glottal flows were computed using the QCP inverse filtering method (Airaksinen et al., 2014). For better comparison, speech and glottal waveforms were resampled to constant length (300 samples). From the figure, it can be seen that in addition to the differences in speech waveforms, differences exist in the shape of glottal source waveforms between the intelligibility categories. Therefore, the estimated glottal source waveforms carry information about the acoustical phenomena that take place at the level of the vocal folds and this information might help in assessing the level of intelligibility in speakers with dysarthria using the speech signal.

3. Glottal inverse filtering and glottal feature extraction from coded speech

In order to extract parameters from the glottal source, the time-domain glottal flow waveform must be first estimated from the speech signal using a GIF method. The existing GIF methods are almost exclusively used with clean speech signals recorded

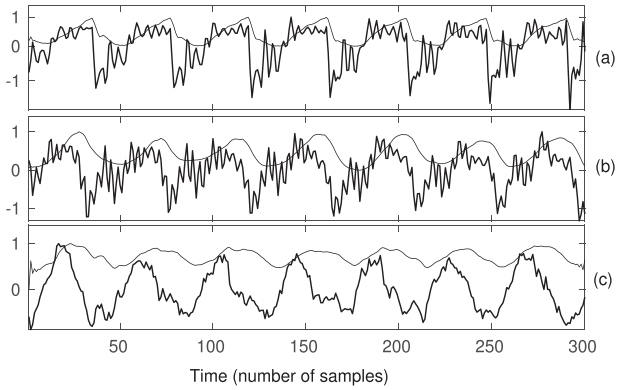

Fig. 1. Speech signals (thick lines) and the corresponding glottal flows (thin lines) computed from parts of the vowel [a] produced by dysarthric speakers of three intelligibility levels: high (a), mediocre (b), and low (c). The speech signals are clean (i.e., not coded telephone speech) and the glottal flows were computed using the QCP inverse filtering method Airaksinen et al. (2014).

in laboratory conditions. Several studies have indicated that the existing GIF techniques are vulnerable even to slight degradation of the input speech signal (such as low-frequency bias caused by breath bursts on the microphone, ambient noise, phase distortion caused by the recording process and inaccurate A/D conversion) (Holmes, 1975; Wong et al., 1979; Narendra et al., 2017; Drugman et al., 2012; Airaksinen et al., 2015). During coding, which is a key task in speech transmission (3GPP TS 26.090, version 10.1.0, 2011), speech is subject to different operations such as band-pass filtering and low bit-rate speech compression. These operations lead to amplitude and phase distortion of speech and to generation of quantization noise that is added to the speech signal (Guillemin et al., 2005; Ireland et al., 2015). As a result, considerable differences are observable in the shape of the glottal pulse when glottal flows estimated using existing GIF methods are compared between clean and coded input speech signals. This phenomenon is illustrated in Fig. 2 using QCP as an example of existing GIF methods. In this article, the glottal source is estimated from coded speech using two GIF methods: (1) QCP which was demonstrated in Airaksinen et al. (2014) to be the most accurate GIF method under clean speech conditions in an evaluation against three widely used GIF techniques (Wong et al. (1979), Alku (1992), Drugman et al. (2009)) and (2) the deep neural net-based glottal inverse filtering (DNN-GIF) method, which is a recently developed robust GIF technique that was shown in Narendra et al. (2019) to have good performance for coded speech compared to four (Wong et al. (1979); Alku (1992); Drugman et al. (2012); Airaksinen et al. (2014)) existing GIF methods.

3.1. GIF methods

QCP (Airaksinen et al., 2014) is a GIF method that is based on one of the most widely used glottal source estimation methods: closed phase analysis (CP) (Wong et al., 1979). Closed phase analysis computes the vocal tract model using just a few data samples that are located in the closed phase of the glottal cycle. The QCP method in turn takes advantage of all data samples of the analysis frame by computing weighted linear prediction (WLP) with a specific attenuated main excitation (AME) weighting function. The AME function downgrades the effect of the glottal source on the vocal tract model in the vicinity of glottal closure instants (GCIs). Using the AME function in the computation of WLP helps to remove the contribution of the (quasi-) open phase and therefore results in good estimates of the vocal tract transfer function. In Airaksinen et al. (2014), the performance of QCP was shown to be 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 the recent study in Narendra et al. (2019) illustrates that the performance of traditional glottal source estimation methods, including QCP, is degraded due to coding, the primary rationale for

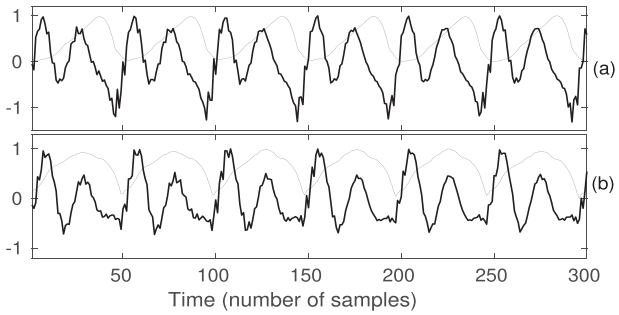

Fig. 2. Illustration of the effects of coding on an existing GIF method. (a) Clean speech signal (thick line) and the corresponding glottal flow (thin line) computed using QCP Airaksinen et al. (2014), and (b) the coded version (thick line) of the speech signal shown in (a) and the corresponding glottal flow (thin line) computed using QCP. The speech signal is extracted from a part of the vowel [a].

using QCP in this work is to analyse how the intelligibility level assessment of dysarthric speech is affected when a conventional GIF method is used to compute glottal features under coded condition.

DNN-GIF (Narendra et al., 2019) is a data-driven method for estimating the glottal flow from coded speech signal. During training, both clean speech signals and the corresponding coded speech signals are utilized. A DNN is used to establish a mapping between spectral features (e.g. LSFs) obtained from the coded speech signal and the glottal flow waveform estimated from the corresponding clean speech by using the QCP method. DNN-GIF utilizes the trained DNNs to map the acoustical features of the coded telephone speech to the time-domain waveform of the glottal flow. The evaluation results in Narendra et al. (2019) demonstrate that 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). This study takes advantage of the same DNN, which was trained as described in Narendra et al. (2019), using both narrowband- and wideband-coded speech utterances from sustained long vowels. The reason for using the same DNN is that the estimated glottal flow waveform is a simple and elementary signal, which is produced by vocal fold vibration (i.e., vocal tract resonances are not present), and utilizing different sets of data for training and testing has minimal influence on the accuracy of GIF (Narendra et al., 2019).

3.2. Glottal feature extraction

In this work, parameterization of glottal flow waveforms obtained by GIF is performed using two types of feature sets. The first one characterizes the glottal flow waveform using a standard set of time- and frequency-domain glottal features. The second feature set utilizes principal component analysis (PCA) to effectively represent every cycle of the glottal flow.

3.2.1. Time- and frequency-domain glottal features (Glottal-TF)

In previous voice production studies, glottal flow waveforms have been typically parameterized using both time-domain and frequency-domain features (for review, see Alku (2011)). These conventional time- and frequency-domain glottal source features, dubbed Glottal-TF in this article, are extracted in the current study using the APARAT Toolbox (Airas et al., 2005). Table 1 summarizes the features used in the Glottal-TF set. From the individual glottal features, H1H2 and HRF are expressed in the dB scale and all other features are computed on a linear scale. The features in Glottal-TF are extracted from voiced speech in 30-ms frames. HRF and H1H2 are calculated pitch-asynchronously once per frame, whereas the remaining features are calculated pitch-synchronously for every glottal cycle after which the features are averaged over the frame. By considering features obtained from all voiced frames of an utterance, the glottal feature vector is created. Eight statistical measures (median, mean, range, maximum, minimum, standard deviation, kurtosis, and skewness) are computed from the glottal feature vector, as well as from its delta vector. This leads to the Glottal-TF feature set consisting of $(12 + 12) \times 8 = 192$ features.

3.2.2. PCA-based glottal features (Glottal-PCA)

PCA-based glottal features (dubbed Glottal-PCA in this article) represent the entire glottal flow waveform with a relatively small number of coefficients using PCA. Parameterization of the glottal flow waveform using PCA was first reported in Gudnason et al. (2009). PCA typically performs an orthogonal transformation of a set of observations into a set of linearly uncorrelated variables called principal components (PCs). PCA-based parameterization of the glottal source has been utilized in speech technology areas such as speech synthesis (Raitio et al., 2013; Narendra and Rao, 2016) and speaker recognition (Drugman and Dutoit,

Table 1
Glottal-TF feature set. For more details, see Airas et al. (2005).

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

2012). However, in the context of assessing intelligibility level in speakers with dysarthria, the PCA-based glottal feature extraction approach has not been used before.

In order to perform PCA-based parameterization, first, the principal components are computed using glottal flow waveforms estimated from sustained vowel utterances. In this work, the same set of sustained vowels that is used in the DNN-GIF training (described in Section 3.1) is utilized and glottal flows are computed using QCP. The glottal flow derivatives are decomposed into smaller segments which are typically two-pitch-period-long and GCI-centered. These segments are then windowed with the Hann window and interpolated to a constant length, and normalized in energy. The glottal segments extracted from every utterance are normalized by subtracting their global mean. Using the normalized glottal segments, the principal component analysis is carried out to obtain eigenvectors (or principal components) and eigenvalues.

PCA-based glottal features are extracted from the flow waveforms using the principal components. First, using dysarthric speech, the estimation of glottal flow waveforms is carried out separately with QCP and DNN-GIF. Following similar steps as explained above, decomposition of glottal flow waveform into two-pitch-period-long glottal segments is performed. Projection of each of these glottal segment on an orthonormal basis formed by the principal components lead to PC weights. This study uses 30 PC weights for the parameterization of glottal segments. The glottal features (30 PC weights) extracted from all glottal cycles of a voiced frame are averaged. The glottal features obtained from all voiced frames of the speech utterance form a glottal feature vector. Eight statistical measures (as mentioned in Section 3.2.1) are calculated from the glottal feature vector, as well as from its delta, leading to the Glottal-PCA feature set consisting of $(30 + 30) \times 8 = 480$ features

In summary, the current study takes advantage of two GIF methods (QCP and DNN-GIF described in Section 3.1) in the estimation of the glottal flow. Both of the resulting waveforms are parameterized with two glottal feature sets (Glottal-TF and Glottal-PCA, described in Section 3.2). In addition, both of the GIF methods and both of the glottal feature sets are combined with the acoustic openSMILE features to build different intelligibility level assessment systems as explained in the next section.

4. Proposed intelligibility level assessment system

4.1. System structure

In order to automatically assess intelligibility in speakers with dysarthria, an intelligibility level assessment system shown in Fig. 3 was developed. The system consists of two main parts: feature extraction and multiclass-SVM. The training phase of the latter is shown in Fig. 4. In training the assessment system, the intelligibility level of every dysarthric speaker in a given training database is expected to be known based on subjective intelligibility tests. In the current study, the UA-Speech database (Kim et al., 2008) is used in training and there are four intelligibility levels (very low, low, mediocre, high) that will be described in Section 5.1. All the speech utterances of a particular speaker are labeled with the corresponding intelligibility level. Both the coded speech signal and the corresponding intelligibility level label are utilized during the system training.

In the feature extraction stage of Fig. 3, coded telephone speech is utilized to compute both acoustic and glottal features. Two sets of acoustic features (openSMILE-1 and openSMILE-2, described in Section 4.2) are computed as reference features. Using glottal flow waveforms that are extracted using the QCP and DNN-GIF methods, two glottal feature sets (Glottal-TF and Glottal-PCA, described in Section 3.2) are extracted and, hence, totally four types of glottal features are computed. In order to identify the intelligibility levels of speakers with dysarthria, a multiclass-SVM using the "one-against-one" method is considered. The multiclass-SVM consists of a number of binary classifiers which are used to distinguish classes C_i and C_j, such that $0 < i < \mathrm{I}$ and $0 < j < i$, where I is the number of intelligibility levels and $(\mathrm{I} \times (\mathrm{I} - 1))/2$ binary SVMs are needed to classify I levels (Fleury et al.,

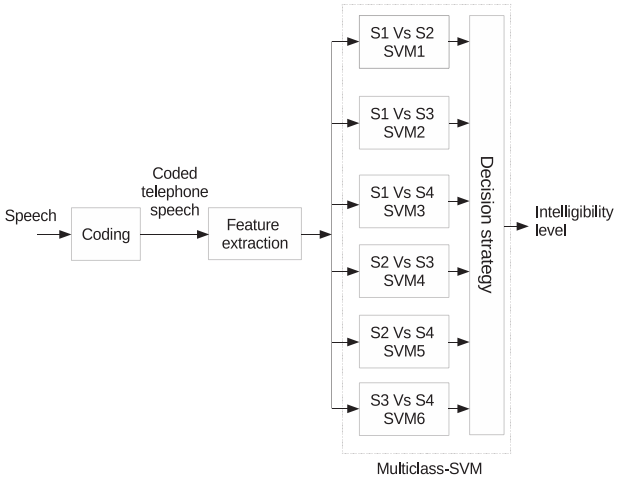

Fig. 3. The proposed intelligibility level assessment method.

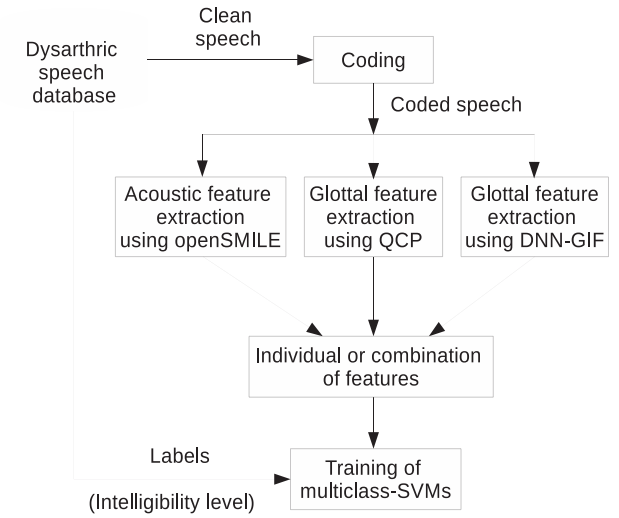

Fig. 4. The training phase of the proposed intelligibility level assessment method.

2010). In Fig. 3, the number of intelligibility levels is I=4, hence the multiclass-SVM includes six binary SVMs (SVM1, SVM2, SVM3, SVM4, SVM5 and SVM6). The decision strategy module outputs intelligibility levels based on the majority voting (best candidate) by considering the output of all binary classifiers. In this approach, multiclass-SVM classifiers are developed using different combination of acoustic and glottal features, computed from coded speech data, to estimate the intelligibility level. This study takes advantage of SVMs in the classification task because they have been widely utilized in the area of pathological speech classification. SVMs have shown good performance, even for smaller volumes of speech data, in comparison to deep neural nets, which require larger volumes of speech data for appropriate training (Kim et al., 2015; Orozco-Arroyave et al., 2014). After the training of SVM classifiers is completed, the system is capable of mapping the input feature set to the target intelligibility level label.

Fig. 4 shows the training phase of the proposed intelligibility level assessment system. First, in order to train the multiclass-SVMs, a dysarthric speech database is utilized (the database used in the current study is described in Section 5.1). The speech utterances of all speakers of the considered database are coded (as explained in Section 4.3). From every coded speech utterance, two sets of baseline acoustic features (openSMILE-1 and openSMILE-2) are first computed. Using QCP and DNN-GIF, glottal flow waveforms are estimated from coded speech and these estimated waveforms are parameterized with the Glottal-TF and Glottal-PCA feature sets. Using features extracted from coded telephone speech utterances and the corresponding intelligibility level labels, multiclass-SVMs are trained. Separate multiclass-SVMs are trained using the openSMILE features, glottal features, and combinations of these features.

After the completion of the training phase, the proposed system can be used to assess the speech intelligibility level in speakers with dysarthria. The features computed from the coded speech utterance are fed to the multiclass-SVM, which outputs the predicted intelligibility label.

4.2. Feature extraction with open SMILE

This study utilizes the openSMILE toolkit (Eyben et al., 2013) for computing acoustic reference features from coded speech. The openSMILE toolkit has been used widely in feature extraction of speech and the openSMILE features have been taken advantage of in different paralinguistic challenges as baselines from INTERSPEECH 2009 (Schuller et al., 2009). Some examples of paralinguistic challenges are recognition of emotions, speaker traits and states, as well as speech pathologies (Schuller et al., 2009; 2012; 2015). The acoustic features computed using the openSMILE toolkit are primarily related to vocal tract spectrum, prosody and voice quality. In this study, two sets of acoustic features (referred to as openSMILE-1 and openSMILE-2) extracted using the openSMILE toolkit are utilized to assess the speech intelligibility level in dysarthria. The openSMILE-1 feature set includes 384 features which were developed in the INTERSPEECH 2009 Emotion Challenge (Schuller et al., 2009). This feature set contains a set of 16 basic acoustic features such as zero-crossing rate, root mean square (RMS) energy, voicing probability, fundamental frequency, and MFCCs. Using the set of 16 acoustic features, along with their derivatives computed from every frame of a speech utterance, an acoustic feature vector is formed. Using the acoustic feature vector of every utterance, 12 statistical functionals (given in Table 2) are calculated to obtain $(16 + 16) \times 12 = 384$ features constituting the final openSMILE-1 feature set. The description of the acoustic features and statistical functionals are given in Table 2.

The openSMILE-2 feature set consists of 6552 features and it is regarded as one of the largest feature sets available in the openSMILE toolkit. Table 2 provides a list of 56 acoustic features that are extracted from each frame of the speech utterance to be parameterized. Fifty-six acoustic features, along with their first and second order derivatives computed from every frame of the

Table 2
Two sets of acoustic features extracted with the openSMILE toolkit. For more details, see Eyben et al. (2013).

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

speech utterance, form the acoustic feature vector. Thirty-nine statistical functionals (given in Table 2) applied to the acoustic feature vector of every utterance form $(56 + 56 + 56) \times 39 = 6552$ features constituting the final openSMILE-2 feature set.

4.3. Coding

Every utterance of the dysarthric speech database is coded using adaptive multi-rate (AMR) coding, which is a widely used speech compression method standardized by the European Telecommunications Standards Institute (ETSI) (3GPP TS 26.090, version 10.1.0, 2011), (Järvinen, 2000). AMR codecs considered in this work operate on two transmission bandwidths - narrowband (300 Hz − 3.4 kHz) and wideband (50 Hz − 7 kHz). Depending on its operational bandwidth, the AMR can be classified as either the AMR narrowband (NB) codec or the AMR wideband (WB) codec. The AMR-NB and AMR-WB codecs operate on sampling frequencies of 8 kHz and 16 kHz respectively. In this work, the openSMILE and glottal feature sets are extracted separately from both NB- and WB-coded speech. By utilizing the features obtained from AMR-coded speech as well as the intelligibility level labels, multiclass-SVMs are trained separately for speech signals coded by the two codecs.

5. Experiments

In this study, experiments were conducted to investigate the effectiveness of glottal features, computed using QCP and DNN-GIF, in the assessment of intelligibility level in speakers with dysarthria using speech signals of these speakers in coded condition. The results of intelligibility level assessment obtained with different classifiers are analyzed. In addition, statistical tests are performed on classifiers developed using the baseline openSMILE features and the combination of the glottal and openSMILE features.

5.1. The UA-Speech database

To develop the intelligibility level assessment system, the Universal access speech (UA-Speech) (Kim et al., 2008) database was utilized. UA-Speech is made freely available for experiments by University of Illinois. The database contains speech samples collected from dysarthric speakers diagnosed with cerebral palsy. The age range of the dysarthric speakers is from 18 to 58. In collecting the data, every speaker uttered isolated words in three blocks (B1, B2, and B3). Each of the three blocks contained 255 words and each block contained a common set of 155 words and 100 uncommon words which vary across the blocks. The 155 words consisted of 10 digits (0 to 9), 26 radio alphabet letters (e.g. 'Alpha', 'Bravo'), 19 computer commands (e.g. "enter", "tab'), and the 100 most common words in the Brown corpus of written English (e.g. "to", "and'). One hundred uncommon words that were uttered in every block were chosen from children's novels digitized by Project Gutenberg (Kim et al., 2008). The UA-Speech data was collected at a sampling frequency of 16 kHz using an eight-microphone array, and spacing between each microphone was 1.5 in. In this experiment, speech data collected from microphone no. 6 of the array was used.

In the UA-Speech database, intelligibility ratings of the speakers, measured using subjective intelligibility tests, are also available. In the subjective tests, five naive listeners were instructed to provide orthographic transcriptions of 225 randomly selected words produced by the dysarthric speakers. Using each listener's transcriptions, the percentage of correct responses was

Table 3
Distribution of the dysarthric speakers in the UA-Speech database according to their speech intelligibility level. (F** for female speakers, M** for male speakers).

Intelligibility level	Speaker ID
High	F05, M08, M10, M14
Mediocre	F04, M05, M11
Low	F02,M07, M16
Very low	F03, M04, M12

calculated. Percentages by the five listeners were then averaged to compute the final intelligibility rating of each speaker. (For more details of the database and subjective tests, please refer to Kim et al. (2008).) Using these intelligibility ratings, the dysarthric speakers of the UA-Speech database were divided in the current study to four intelligibility level categories as shown in Table 3. The intelligibility ratings of all speakers were in the range of 1−100. Speakers with ratings in the range of 1−25 were categorized in terms of their speech intelligibility level as 'very low", those in the range of 26−50 as 'low", those in the range of 51−75 as 'mediocre' and those in the range of 76−100 as 'high'.

It is worth pointing out that the intelligibility assessment scenario addressed in the current investigation deviates from typical speech intelligibility studies in engineering where impacts of issues such as coding, communication channel and additive noise on intelligibility are studied (Van Kuyk et al., 2018; Jokinen et al., 2012; Tang and Cooke, 2010; Sauert and Vary, 2006; Tang and Cooke, 2011). While intelligibility in these studies is determined using either subjective evaluations or instrumental measures separately for each linguistic unit (e.g. word or sentence) of the evaluation data, the current study focuses on the *overall* intelligibility level of utterances spoken by dysarthric speakers. To investigate the overall intelligibility level in dysarthric speakers, the current study used 765 speech utterances from every speaker. In addition, it is worth pointing out that this study investigates a relatively small number of speakers (13) for assessing the intelligibility level in patients with dysarthria. This relatively small number is in contrast to the number of speakers in other areas of speech technology, such as speech recognition (Hsu et al., 2019) and speaker verification (Kinnunen et al., 2006), where it is possible to collect speech from healthy speakers by recording utterances from hundreds or even thousands of talkers. It should be noted, however, that in dysarthric speech assessment, a similar small number of speakers (less than 15) has been used also in previous studies (Narendra and Alku, 2018; Paja and Falk, 2012; Bhat et al., 2017).

5.2. Experimental setup

In the intelligibility level assessment system, every speech utterance of the UA-Speech database is coded using the two AMR codecs described in Section 4.3. Both the openSMILE and glottal feature sets are extracted separately from NB- and WB-coded speech utterances. In this work, the Glottal-TF feature set obtained using the QCP and DNN-GIF methods are dubbed 'Glottal-TF (QCP)' and 'Glottal-TF (DNN-GIF)' respectively. The Glottal-PCA feature set computed using the QCP and DNN-GIF methods are dubbed 'Glottal-PCA (QCP)' and 'Glottal-PCA (DNN-GIF)' respectively. Both acoustic and glottal feature sets are individually normalized by subtracting the global mean and dividing by the global standard deviation. The coded speech utterances are analyzed in 30-ms frames using a 15-ms frame shift. From every speaker of the UA-Speech database, 90% of his/her speech data is used for training and the remaining 10% is used for validation. The training data is primarily utilized for developing classification models and for computing their accuracy. Validation data is utilized in experiments for selecting optimal features from the entire set of features and for choosing the suitable hyper parameters. In this work, the selection of the optimal features is performed using the sequential forward feature selection (SFFS) (Reunanen, 2003) algorithm. SFFS constructs candidate feature subsets by incrementally adding each of the features, and each candidate feature subset is evaluated by computing the classification accuracy with the 10-fold cross-validation strategy on the validation data. The best candidate feature set that leads to the highest classification accuracy is selected. This selection of the optimal features results both in improved computational efficiency and in better generalization of the system. Using feature sets obtained before and after applying SFFS, multiclass-SVM classifiers are trained separately for both versions of coded speech. In the multiclass-SVM classifier structure, the Gaussian, radial basis function is used as a kernel in every binary SVM classifier. Two hyper parameters (kernel parameter γ and penalty parameter C) required for training of a binary SVM classifier are optimally chosen by conducting a grid search in which the parameters were varied in the range of 10^{-3}–10^{3} in multiples of 10. The pair (C, γ) that led to the highest intelligibility level accuracy is selected as the most appropriate. For computing intelligibility level accuracy, the leave-one-speaker-out cross validation strategy is used on validation data.

In order to compute the classification accuracy of a multiclass-SVM classifier trained with a particular feature set, the leave-one-speaker-out cross validation strategy is used. In this strategy, all speech utterances of one speaker (100% of data) are kept out for testing and the speech data of the remaining speakers are used for training (90% of every speaker's data) and validation (10% of every speaker's data). The validation data is used to optimize the classification models, i.e., to select suitable hyper parameters and to reduce the size of the feature set. The training data is used to develop the classification model. The held-out speakerâs data or test data is not used at any stages in the development of the classification model and it is totally unseen to the classification models. The held-out data is used only to compute the classification accuracy. This process is repeated with each speaker being held out exactly once for testing. The classification accuracies obtained from all held-out speakers' data are averaged to obtain the final accuracy. For a given speech utterance, the intelligibility level predicted from a multiclass-SVM classifier is compared to the known intelligibility level label of the speaker. The classification accuracy is calculated as the ratio of the number of correctly classified speech utterances to the total number of the classified speech utterances of the speaker.

After computing the classification accuracies, statistical analyses were conducted using Cochran's Q test (Daniel, 1978) to compare the classifiers developed with different feature sets. Cochran's Q test is known to be a more general form of McNemar's test that is best suitable for comparison of multiple classifiers (Narendra and Alku, 2019). Cochran's Q test is computed individually to all classifiers and separately for both NB- and WB-coded speech.

Table 4
Accuracy in classifying the intelligibility level (with four classes) in the dysarthric speakers of the UA-Speech database. Accuracy is shown for different features sets both in NB-coded and in WB-coded speech using both the non-reduced and reduced feature sets.

	Classification Accuracy	
Feature set (NB-coded)	Without feature selection (%)	With feature selection (%)
openSMILE-1	35.13	43.02
openSMILE-2	47.76	56.23
Glottal-TF (QCP)	45.91	46.13
Glottal-PCA (QCP)	42.64	44.11
Glottal-TF (DNN-GIF)	43.35	46.82
Glottal-PCA (DNN-GIF)	39.99	42.28
openSMILE-1 + Glottal-TF (QCP)	58.71	58.81
openSMILE-2 + Glottal-TF (QCP)	59.01	59.22
openSMILE-1 + Glottal-PCA (QCP)	44.64	47.28
openSMILE-2 + Glottal-PCA (QCP)	46.25	56.83
openSMILE-1 + Glottal-TF (DNN-GIF)	47.43	59.27
openSMILE-2 + Glottal-TF (DNN-GIF)	49.87	60.25
openSMILE-1 + Glottal-PCA (DNN-GIF)	43.19	49.10
openSMILE-2 + Glottal-PCA (DNN-GIF)	45.82	57.08
Feature set (WB-coded)	Without feature selection (%)	With feature selection (%)
openSMILE-1	38.85	44.41
openSMILE-2	48.89	59.75
Glottal-TF (QCP)	49.31	51.01
Glottal-PCA (QCP)	36.53	41.75
Glottal-TF (DNN-GIF)	48.10	51.93
Glottal-PCA (DNN-GIF)	29.18	44.47
openSMILE-1 + Glottal-TF (QCP)	57.56	62.73
openSMILE-2 + Glottal-TF (QCP)	68.90	67.72
openSMILE-1 + Glottal-PCA (QCP)	46.70	49.43
openSMILE-2 + Glottal-PCA (QCP)	55.58	60.75
openSMILE-1 + Glottal-TF (DNN-GIF)	47.66	68.83
openSMILE-2 + Glottal-TF (DNN-GIF)	53.95	69.55
openSMILE-1 + Glottal-PCA (DNN-GIF)	38.27	48.80
openSMILE-2 + Glottal-PCA (DNN-GIF)	47.39	60.09

5.3. Results

During testing, a given input speech utterance is classified into one of the four intelligibility levels based on the output of the multiclass-SVM classifier. The four-class classification accuracies obtained using the reduced and non-reduced feature sets for the two versions of coded speech are shown in Table 4. From the table, it can be observed that among the different feature sets, the performance of the classifiers developed using the reduced feature sets is higher than that of the classifiers using the non-reduced feature sets. In comparing the openSMILE and glottal feature sets, the openSMILE-2 feature set shows the highest accuracy for both versions of coded speech.

By analysing classifier accuracies obtained using the Glottal-TF and Glottal-PCA feature sets, it can be seen that the accuracy achieved with these two feature sets varies in the range of $40 - 50\%$. Since the studied scenario is a four-class classification, the results indicate the importance of glottal features for the assessment of speech intelligibility level in speakers with dysarthria. In comparing the systems developed using the two GIF methods, the classifiers developed with DNN-GIF show slightly better accuracies for both versions of coded speech. From the two glottal feature sets, the Glottal-TF feature set shows improved accuracy compared to the Glottal-PCA feature set in all cases.

From Table 4, it can be observed that the performance of the classifiers developed with the combination of glottal features and the openSMILE features is better than the performance of the classifiers where either of these sets were used alone. The classifier trained by combining the openSMILE-2 feature set to the Glottal-TF (DNN-GIF) feature set results in the highest accuracy among all other combinations for both NB-coded and WB-coded speech.

Cochran's Q test was carried out on classifiers developed with various feature sets (baseline features and combination of baseline and glottal features) obtained from the two versions of coded speech. The Cochran's Q tests rejected null hypothesis at α = 0.0056 (*degrees of freedom* = 9) which indicates that there are differences in classifiers. In addition, a pairwise Cochran's Q test was performed between the classifier that resulted in the highest accuracy and all other classifiers separately for both narrowband and wideband coded speech. For both versions of coded speech, the pairwise test was carried out between the classifiers trained using the combination of the openSMILE-2 and Glottal-TF (DNN-GIF) feature sets (which is regarded as the best

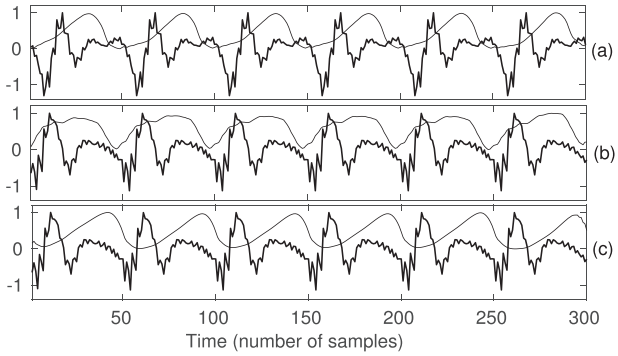

Fig. 5. (a) Clean speech signal (thick line) and the corresponding glottal flow (thin line) computed using QCP Airaksinen et al. (2014). WB-coded version (thick line) of the speech signal shown in (a) and the corresponding glottal flow (thin line) computed using (b) QCP and (c) DNN-GIF. The speech signal is extracted from a part of the vowel [a] .

performing classifier) and all other classifiers. The results indicate that the best performing classifier was significantly ($p < 0.005$) better than all other classifiers, except the classifier trained with the openSMILE-1 + Glottal-TF (DNN-GIF) feature set.

In order to better understand the differences in the classification results between the DNN-GIF- based glottal features and the QCP-based glottal features, an additional study was conducted to demonstrate glottal flows estimated by QCP and DNN-GIF. Fig. 5 shows a reference glottal flow waveform (panel (a)) obtained from a clean utterance using QCP. The glottal flow estimated from the corresponding WB-coded version of the utterance is shown in panel (b) when the estimation was computed with QCP and in panel (c) when the estimation was computed using DNN-GIF. It can be clearly observed from the two glottal flows computed using WB-coded speech that the one estimated with DNN-GIF (panel (c)) is much closer to the reference glottal flow (panel

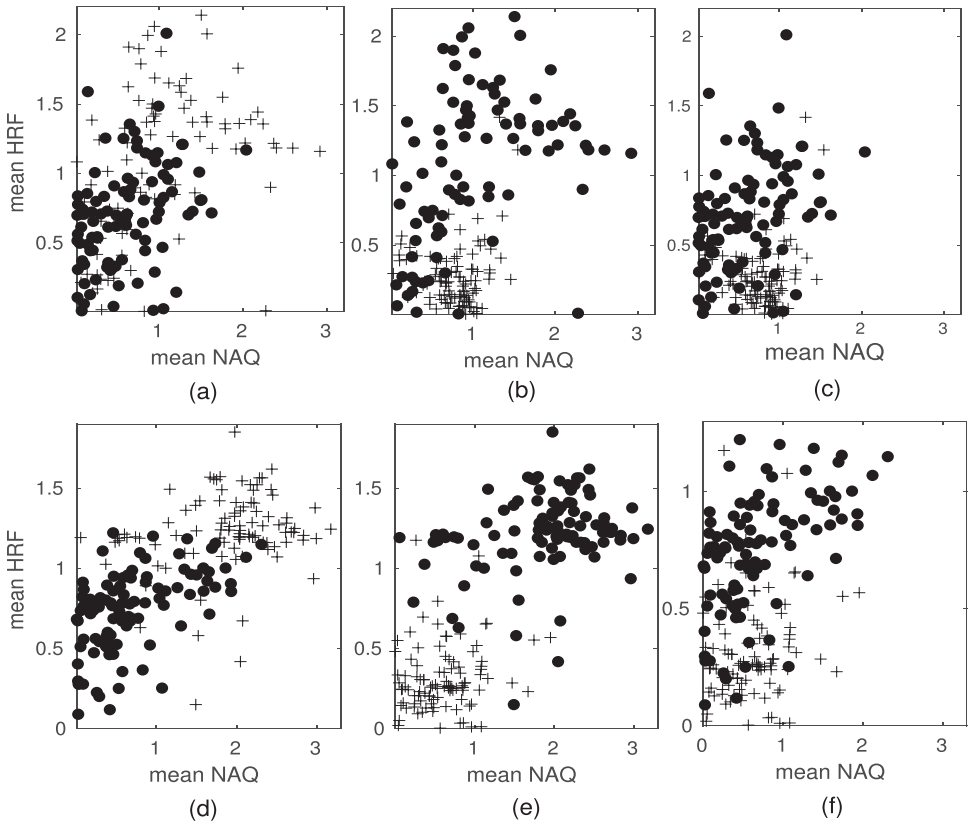

Fig. 6. Scatter plots between two glottal features (mean NAQ and mean HRF) for dysarthric speakers of different intelligibility levels in the UA-Speech database. The glottal features were computed using WB-coded speech. The estimation of the glottal flow was conducted with QCP (panels (a)-(c)) and DNN-GIF (panels (d)-(f)). Three intelligibility level pairs are shown: high (circles) vs. mediocre (crosses) in panels (a) and (d), mediocre (circles) vs. low (crosses) in panels (b) and (e), and high (circles) vs. low (crosses) in panels (c) and (f).

(a)) than the one estimated using QCP (panel (b)). More detailed objective evaluation of the QCP and DNN-GIF methods under coded condition is provided in our recent study in Narendra et al. (2019).

Fig. 6 depicts the scatter plots of two features (mean NAQ and mean HRF) of the Glottal-TF set that were estimated using QCP and DNN-GIF for the dysarthric speakers of different intelligibility levels. The features were computed from 100 WB-coded speech utterances in each intelligibility level category. Even though a slight overlap is observed in the scatter plots, speech utterances of different intelligibility levels can be separated with appropriate decision boundaries. On pairwise comparison of the scatter plots between QCP and DNN-GIF (i.e. (a) vs. (d), (b) vs. (e), and (c) vs. (f)), it can be observed that the scatter plots of QCP are distributed more widely compared to those of DNN-GIF. In addition, it can be seen that the overlap in scatter plots between intelligibility level pairs is slightly larger for QCP (clearly visible particularly in (a)) compared to DNN-GIF. In conclusion, the data shown in Figs. 5 and 6 demonstrate the effectiveness of DNN-GIF in comparison to QCP when glottal flows and their features are computed from WB-coded speech. The observations made between DNN-GIF and QCP in this section are in line with the objective results obtained in the classification experiments for WB-coded speech (Table 4) that show better performance for DNN-GIF compared to QCP.

The glottal features proposed in this study were shown to be effective in assessing the intelligibility levels of speakers with dysarthria. It is, however, worth pointing out that these glottal features are generic parameters depicting the functioning of the human voice production mechanism and they were not designed specifically for dysarthric speech. Therefore, the glottal features used in this study can also also be taken advantage of in investigating other speech disorders. In addition, the glottal features can be potentially be also used in other tasks such as in analysis of voice quality and in recognition of different paralinguistic cues (e.g., emotion, gender, and age) from speech.

6. Conclusion

An automatic intelligibility level assessment system was proposed for coded telephone speech of dysarthric speakers based on glottal features. In the proposed method, multiclass-SVMs were trained to predict the intelligibility level in speakers with dysarthria by utilizing different sets of acoustic and glottal features obtained from coded speech data. In order to extract acoustic feature sets, the openSMILE toolkit was utilized and two sets of glottal features were computed from the glottal flow waveform. Estimation of glottal flow waveforms from coded telephone speech (compressed with two versions of AMR codecs) was performed using two GIF methods: QCP and DNN-GIF. Intelligibility level assessment results showed that the utilization of glottal features lead to good classification accuracy for both versions of coded speech of the UA-Speech database. The evaluation results showed that from the two glottal feature sets, the time- and frequency-domain glottal features resulted in better accuracy than the PCA-based glottal features. The results also indicated higher accuracies for the glottal features estimated using DNN-GIF compared to the features obtained from QCP. This illustrates the importance of DNN-GIF in the glottal flow estimation from coded telephone speech. Improvement in classification accuracy was also observed in the experiments when the baseline openSMILE features were combined with the glottal features. The best classification performance was achieved when the Glottal-TF (DNN-GIF) feature set was combined with the openSMILE-2 feature set.

To the best of our knowledge, the present work is the first systematic examination of the automatic assessment of intelligibility level in dysarthric speakers from coded speech. Possible extensions to the present work are as follows: In addition to the AMR codecs, recent codecs (such as the Enhanced Voice Services (EVS) codec (3GPP TS 26.445, 2014)) can be used for the evaluation of intelligibility level assessment. The proposed method based on glottal features can be utilized for the speech-based assessment of neurological motor diseases such as Alzheimer's disease, Parkinson's disease, and ALS. In addition, the approach studied in this work can be investigated in various paralinguistic areas such as in recognition of emotion states and speaker traits.

Acknowledgements

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

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., 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., Alku, P., 2006. Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalised amplitude quotient. Phonetica 63 (1), 26–46.
Airas, M., Alku, P., Vainio, M., 2007. Laryngeal voice quality changes in expression of prominence in continuous speech. In: Proc. International Workshop on Models and Analysis of Vocal Emissions in Biomedical Applications (MAVEBA), pp. 135–138.
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., Airas, M., Björkner, E., Sundberg, J., 2006. An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity. J. Acoust. Soc. Am. 120 (2), 1052–1062.

Alku, P., Vilkman, E., 1996. A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers. Folia Phoniatrica et Logopaedica 48 (5), 240–254.
Arroabarren, I., Carlosena, A., 2006. Effect of the glottal source and the vocal tract on the partials amplitude of vibrato in male voices. J. Acoust. Soc. Am. 119 (4), 2483–2497.
Bhat, C., Vachhani, B., Kopparapu, S.K., 2017. Automatic assessment of dysarthria severity level using audio descriptors. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5070–5074.
Campbell, N., Mokhtari, P., 2003. Voice quality: the 4th prosodic dimension. In: Proc. International Congress of Phonetic Sciences, pp. 2417–2420.
Carmichael, J., 2007. Introducing objective acoustic metrics for the Frenchay Dysarthria Assessment procedure. University of Sheffield.
Chien, Y.-R., Borský, M., Gudnason, J., 2017. Objective severity assessment from disordered voice using estimated glottal airflow. In: Proc. Interspeech, pp. 304–308.
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. International Journal of Language and Communication Disorders 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.
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. Computer Speech and Language 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.
Duffy, J.R., 2000. Motor speech disorders: Clues to neurologic diagnosis. Humana Press, Totowa, NJ.
Duffy, J.R., 2012. Motor speech disorders: Substrates, differential diagnosis, and management, 3 Elsevier Health Sciences.
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.
Fleury, A., Vacher, M., Noury, N., 2010. SVM-Based multimodal classification of activities of daily living in health smart homes: sensors, algorithms, and first experimental results. IEEE Trans. Inf. Technol. Biomed. 14 (2), 274–283.
Fraser, K.C., Rudzicz, F., Hirst, G., 2016. Detecting late-life depression in Alzheimer's disease through analysis of speech and language. In: Proc. Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 1–11.
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.
Goetz, C.G., Stebbins, G.T., Wolff, D., DeLeeuw, W., Bronte-Stewart, H., Elble, R., Hallett, M., Nutt, J., Ramig, L., Sanger, T., Wu, A.D., 2009. Testing objective measures of motor impairment in early Parkinson's disease: feasibility study of an at-home testing device. Movement Disorders 24 (4), 551–556.
Gudnason, J., Thomas, M.R.P., Naylor, P.A., Ellis, D.P.W., 2009. Voice source waveform analysis and synthesis using principal component analysis and gaussian mixture modelling. In: Proc. Interspeech, pp. 108–111.
Guillemin, B.J., Watson, C.I., Dowler, S., 2005. Impact of the GSM AMR speech codec on acoustic parameters used in forensic speaker identification. In: Proc. International Symposium on DSP and Communications Systems (DSPCS'2005), pp. 61–66.
Hanratty, J., Deegan, C., Walsh, M., Kirkpatrick, B., 2016. Analysis of glottal source parameters in Parkinsonian speech. In: Proc. IEEE International Conference of the Engineering in Medicine and Biology Society (EMBC), pp. 3666–3669.
Holmes, J.N., 1975. Low-frequency phase distortion of speech recordings. J. Acoust. Soc. Am. 58 (3), 747–749.
Hsu, W.-N., Harwath, D., Glass, J., 2019. Transfer learning from audio-visual grounding to speech recognition. In: Proc. Interspeech, pp. 3242–3246.
Ireland, D., Knuepffer, C., McBride, S.J., 2015. Adaptive multi-rate compression effects on vowel analysis. Front Bioeng Biotechnol 3, 1–9.
Järvinen, K., 2000. Standardisation of the adaptive multi-rate codec. In: Proc. European Signal Processing Conference (EUSIPCO).
Jokinen, E., Yrttiaho, S., Pulakka, H., Vainio, M., Alku, P., 2012. Signal-to-noise ratio adaptive post-filtering method for intelligibility enhancement of telephone speech. J. Acoust. Soc. Am. 132, 3990–4001.
Kadi, K.L., Selouani, S.A., Boudraa, B., Boudraa, M., 2014. Automated diagnosis and assessment of dysarthric speech using relevant prosodic features. Transactions on Engineering Technologies, Springer 529–540.
Kadi, K.L., Selouani, S.A., Boudraa, B., Boudraa, M., 2016. Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge. Biocybernetics and Biomedical Engineering 36, 233–247.
Kent, R.D., Weismer, G., Kent, J.F., Rosenbek, J.C., 1989. Toward phonetic intelligibility testing in dysarthria. Journal of Speech Hearing Disorders 54 (4), 482–499.
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. Computer Speech and Language 29, 132–144.
Kim, M.J., Kim, H., 2012. Combination of multiple speech dimensions for automatic assessment of dysarthric speech intelligibility. In: Proc. Interspeech, pp. 1323–1326.
Kinnunen, T., Alku, P., 2009. On separating glottal source and vocal tract information in telephony speaker verification. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 4545–4548.
Kinnunen, T., Karpov, E., Fränti, P., 2006. Real-time speaker identification and verification. IEEE Trans Audio Speech Lang Process 14 (1), 277–288.
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.
Laaridh, I., Kheder, W.B., Fredouille, C., Meunier, C., 2017. Automatic prediction of speech evaluation metrics for dysarthric speech. In: Proc. Interspeech, pp. 1834–1838.
Lansford, K.L., Liss, J.M., Norton, R.E., 2014. Free-classification of perceptually similar speakers with dysarthria. Journal of Speech, Language, and Hearing Research 57, 2051–2064.
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.
Martínez, D., Green, P., Christensen, H., 2013. Dysarthria intelligibility assessment in a factor analysis total variability space. In: Proc. Interspeech, pp. 2133–2137.
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., 2019. Estimation of the glottal source from coded telephone speech using deep neural networks. Speech Commun 106, 95–104.
Narendra, N.P., Alku, P., 2018. Dysarthric speech classification using glottal features computed from non-words, words and sentences. In: Proc. Interspeech, pp. 3403–3407.
Narendra, N.P., Alku, P., 2019. Dysarthric speech classification from coded telephone speech using glottal features. Speech Commun 110, 47–55.

Narendra, N.P., Rao, K.S., 2016. A deterministic plus noise model of excitation signal using principal component analysis for parametric speech synthesis. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5635–5639.
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.
Paja, M.S., Falk, T.H., 2012. Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech. In: Proc. Interspeech, pp. 62–65.
Polikoff, J.B., Bunnel, H.T., 1999. The Nemours database of dysarthric speech: A perceptual analysis. In: Proc. International Congress of Phonetic Sciences, pp. 783–786.
Rabiner, L., Schafer, R., 1978. Digital processing of speech signals. Prentice-Hall signal processing series. Prentice-Hall.
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.
Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., Alku, P., 2011. HMM-Based speech synthesis utilizing glottal inverse filtering. IEEE Trans Audio Speech Lang Process 19 (1), 153–165.
Reunanen, J., 2003. Overfitting in making comparisons between variable selection methods. J. Mac. 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.
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.
Sauert, B., Vary, P., 2006. Near end listening enhancement: Speech intelligibility improvement in noisy environments. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 493–496.
Schuller, B., Steidl, S., Batliner, A., 2009. The INTERSPEECH 2009 Emotion challenge. In: Proc. Interspeech, pp. 312–315.
Schuller, B., Steidl, S., Batliner, A., Hantke, S., Hönig, F., Orozco-Arroyave, J.R., Nöth, E., Zhang, Y., Weninger, F., 2015. The INTERSPEECH 2015 Computational paralinguistics challenge: Nativeness, Parkinson's and Eating condition. In: Proc. Interspeech, pp. 478–482.
Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., Weiss, B., 2012. The INTERSPEECH 2009 Speaker trait challenge. In: Proc. Interspeech, pp. 254–257.
Skodda, S., Grönheit, W., Mancinelli, N., Schlegel, U., 2013. Progression of voice and speech impairment in the course of Parkinson's disease: a longitudinal study. Parkinsons Dis 2013, 1–8.
Tang, Y., Cooke, M., 2010. Energy reallocation strategies for speech enhancement in known noise conditions. In: Proc. Interspeech, pp. 1636–1639.
Tang, Y., Cooke, M., 2011. Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints. In: Proc. Interspeech, pp. 345–348.
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 Kuyk, S., Kleijn, W.B., Hendriks, R.C., 2018. An evaluation of intrusive instrumental intelligibility metrics. IEEE/ACM Trans Audio Speech Lang Process 26 (11), 2153–2166.
Van Nuffelen, G., Middag, C., De Bodt, M.S., Martens, J.-P., 2009. Speech technology-based assessment of phoneme intelligibility in dysarthria. International Journal of Language and Communication Disorders 44 (5), 716–730.
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.