CS 69002: Seminar II - Report

Dysarthria Speech Classification

(Using Glottal Features)

Submitted by **Aditya Anand** (20CS60R24)

Under the Guidance of

Prof. K. Sreenivasa Rao

and

Ph.D. Mentor, Miss Madhu Keerthana

Department of Computer Science and Engineering

Indian Institute of Technology, Kharagpur

Kharagpur-721302, India

ABSTRACT

This report shows the study of Dysarthria Speech classification and various components involved in it like the Databases (TORGO and UN Speech), classifier (SVM), Feature Extractor (openSMILE, GIF and DNN-GIF), Feature selector (SFFS), and Validation Methods (K –Fold Cross validation). Also, it shows the importance of Glottal Features in the enhancement of classification of Dysarthria Speech.

ACKNOWLEDGEMENT

The successful completion of any task can't be complete without acknowledging the people who made it all possible and whose constant guidance and encouragement leads to success. I am grateful to **Prof. K. Sreenivasa Rao** and **Ms. Madhu Keerthana**, Computer Science and Engineering Department, IIT Kharagpur for the support. I owe a debt of gratitude to them for incorporating in me the idea for this seminar report, helping me undertake this seminar and also for being there whenever I needed their assistance. I also place on record, our sense of gratitude to my fellow batch-mates and all, who directly or indirectly have lent a helping hand in this venture. I feel proud and privileged in expressing my deep sense of gratitude to all those who have helped me in presenting this seminar.

Table of Contents

1. Introduction

Dysarthria is a neuro-motor disease that results in nerve damage to the motor parts that are responsible for producing speech. Dysarthria are usually the result of nerve damage (i.e., cerebral palsy, brain tumor, brain injury, and stroke) or a sign of neurodegenerative disease. Dysarthria speech is often associated with a decrease in the volume of vocabulary and language fluctuations, abnormal speech, unintelligible speech, and fluctuating speech - factors that reduce speech comprehension.

Some of the signs and symptoms of Dysarthria Speech are slow and slurred speech, Rapid Speech difficult to understand, inability to whisper or speak louder, uneven speech volume, problems in facial muscles or throat, etc...

2. Motivation for a Data-driven Model

Speech tests can be performed using a traditional method, which involves speech therapists who perform comprehension tests to understand the presence of dysarthria, as well as to demonstrate its severity. Independent cognitive tests, however, are expensive, effective, and often prone to internal disease discrimination due to patients' familiarity and their speech impediments.

This stimulates and motivates the development of an objective method of diagnosing dysarthria speech. Dysarthria speech testing performed two tasks:

- To identify the presence of dysarthria from a given speech signal.
- To estimate the severity of dysarthria.

Both of these activities are important diagnostic steps that help to make clinical decisions about treatment or medication for patients.

3. Dysarthria Speech Databases

To run the data-driven model, data is required in the form of speech databases.

3.1. TORGO Database

To develop the dysarthric speech classification system, the TORGO database was utilized [R1]. This database has speech recording of 7 patients (3 females and 4 males), and speech recording from 7 healthy speakers (3 women and 4 men), aging between 16 to 50 years. Here the speech signals are divided into 3 categories namely non-worlds (like long vowels utterance), words and sentences. Non-words consist of few repetitions of /iyp-ah/, /ah-p-iy/, and /p-ah-t-ah-k-ah/ and high and low pitched vowels maintained over 5 s (e.g., “Say ‘eee’ in a high pitch for 5 s”).

In this study [R1], words were used in all three stages of the TORGO speech signal, recorded with consecutive 16-kHz sampling microphones. In each of the three phases, speech samples of seven patients (three women and four men) with dysarthria and seven healthy speakers (three women and four men) were considered. For word and sentence level speech signals, 80 words from each speaker are used (with the exception of two loudspeakers at the sentence level, only 23 and 28 words are used for lack of access to the recording) and non-verbal words, 8- 9 sentences from each speaker (available in the database) are used for dysarthric speech expression.

3.2. UA- Speech Database

The UA-Speech database contains a collection of speech samples from 15 patients (4 women and 11 men) diagnosed with a brain disability and speech from 13 healthy control speakers (4 women and 9 men). Patients' ages range from 18 to 58. Each person is asked to name 765 words in three blocks and each block has 255 words, including 155 common words in all three blocks and 100 unique words across the blocks. 155 words contain 19 standard computer commands (e.g., 'Insert', 'tab'), 10 digits (0 to 9), and 26 radio characters (eg, 'Alpha', 'Bravo'), as well as the 100 most common names in the Brown corpus of written English (e.g., 'to', 'and'). In this work[R2], Dysarthric speech data was recorded using a set of eight microphones, a sample at 16 kHz and each microphone separated by 1.5 inches.

4. Feature Extraction

Now, once we have the speech utterances from the databases, we now move towards extracting some features that will be useful for training the Data-Driven Model.

4.1. Acoustic Feature Extraction

Acoustic parameters are extracted from speech using openSMILE, a freely available tool. The openSMILE features have been used as a basis for various paralinguistic challenges from INTERSPEECH 2009. The acoustic features released by openSMILE in particular represent spectrum, prosody and voice quality.

In this work, two sets of acoustic features described in the openSMILE toolkit are used for dysarthric speech classification. The first set (referred to in this activity [R1] as openSMILE-1) is the INTERSPEECH 2009 Emotion Challenge has a feature set with 384 features. This feature comprises 16 acoustic features extracted from all frames (*described in Table 1*). A collection of 16 acoustic features and their first order derivatives found in all speech frames forms an acoustic feature vector. The 12 mathematical performances (*shown in Table 1*) is calculated from the acoustic display of the voice to find (16 + 16) × 12 = 384 elements representing the set of openSMILE-1 feature.

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 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 maximum and minimum	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 of peaks, mean distance between peaks, arithmetic, geometric and quadratic means

Table 1: *Two openSMILE featureset.[R1]*

The second set (referred to in this work [R1] as openSMILE-2) is a large set of openSMILE feature with 6552 features. This is the largest feature set according to the number of features in the openSMILE toolkit. The largest feature set is selected to include the acoustic

details that can assist in dysarthric speech separation. A set of 56 acoustic features (given in Table 1) are extracted from all frames. The 56 acoustic features and their first and second order exits create acoustic level features. As in openSMILE-1, mathematical functionality is applied to acoustic feature vectors that are extracted from all speech frames. Instead of 12, 39 mathematical operations (shown in Table 1) are used to find (56 + 56 + 56) × 39 = 6552 features representing the set of openSMILE-2 feature.

4.2. QCP based Glottal Inverse Filtering

QCP [R3] is one of the recently proposed GIF methods for estimating glottal source from speech. The QCP method is based on closed phase analysis (CP) measurement that measures the response of voice fragments using the corresponding predictive covariance method from a few speech samples found in the closed phase of the glottal cycle. It makes use of the Weighted Linear Prediction Model to estimate the glottal wave flow. WLP is quite similar to the conventional LP [R3],

$$s_n = \sum_{i=1}^{p} s_{n-i} a_i + e_n$$

Here s_n: the nth sample of a speech wave,
e_n : the nth sample of an excitation (residual) wave,
a_i : the ith predictor coefficient, and
p : the LP order.

WLP imposes temporal weighting on the square of the residual, thus it differs from conventional LP [R3]:

$$E = \sum_{n=n_1}^{n_2} \left(s_n - \sum_{i=1}^{p} s_{n-i} a_i \right)^2 W_n$$

Here, E : the total residual energy,
$n_1 = 1$ and $n_2 = N + p$ for the autocorrelation case,
N : the frame length (in samples), and
Wn : weighting function.

Therefore, in this work, the glottal flow waveform is estimated from the QCP method. From the forms of glottal flow waveforms, glottal parameters based on time and frequency domain are extracted. APARAT toolbox is used to extract these features (as shown in Table 2).

Time-domain parameters	
OQ1	Open quotient, computed from primary glottal opening
OQ2	Open quotient, computed from 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, computed from primary glottal opening
SQ2	Speed quotient, computed from secondary glottal opening
Frequency-domain parameters	
H12	Difference between first two glottal harmonics
PSP	Parabolic spectrum parameter
HRF	Harmonic richness factor

Table 2: Time and frequency domain Glottal Parameters [R1].

The above 8 statistical measures are computed from the glottal parameter vector as well as from its delta vector: mean, median, min, max, standard deviation, range, skewness and kurtosis. Thus, resulting in (12 + 12) × 8 = 192 parameters representing the glottal feature set.

4.3. Deep Neural Networks based GIF

Standard GIF methods are known to be more sensitive to minor signal distortions of input speech. To better measure the glottal source from coded speech, a deep neural net-based glottal inverse filter (DNN-GIF) has recently been proposed. In DNN-GIF, a deep neural network (DNN), trained in a controlled manner using glottal flow measured from pure speech, maps of signal sensor input directly into the time signal of glottal flow. According to the experiments carried, the accuracy of DNN-GIF is better than that of QCP [R2].

- **Deep Neural Networks:**

Deep learning is a subset of machine learning methods based on artificial neural networks with representative learning. Artificial neural networks (ANNs) were inspired by information processing and

distributed communication nodes in biological systems mainly the brain. It tries to mimic the idea of how the neurons in brain work and how they communicate. A DNN is just an ANN with multiple hidden layers between the input and the output layer.

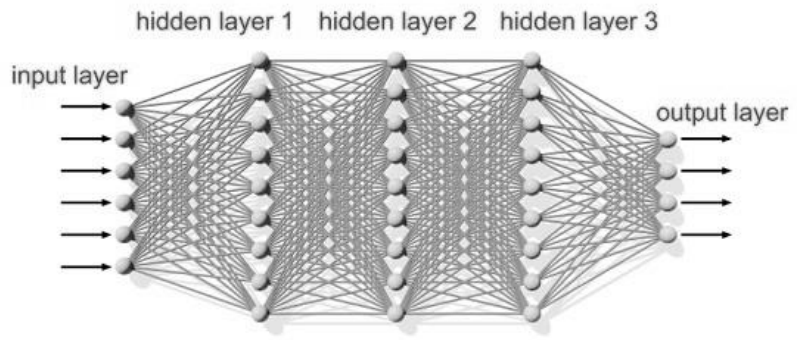

Fig. 1: Basic structure of DNN (Src: Google Images)

As shown in Fig.1, a DNN has an input layer, many hidden layers and an output layer. Each layer has neurons/nodes and nodes in i^{th} layer is only connected to nodes in $(i+1)^{th}$ layer. The input, in the form of vectors, is fed to the nodes in input layer, and each node has its own biases and each edge has its own weights. The values in the nodes are multiplied by the weights of the edges it is connected to and added to its bias, which is then passed over an activation function of that layer to the connected nodes in the next layer. The activation function may be linear or non-linear, generally it is non-linear.

The DNN-GIF method relies on different linear variations of DNNs to measure glottal flow. In this way, DNN is trained to map the spectral features extracted from the coded telephone conversation and the reference forms of reference glottal flow wave from pure speech. Glottal flow waveforms are measured from pure speech using QCP method. Using GIF-rated glottal flow forms, two sets of glottal parameters are extracted from the current study [R2]. The first set of parameters captures time elements- and the frequency elements of the glottal flow waveform. The second parameter set aims to represent the entire glottal flow waveform formula using PCA.

5. Feature Selection

The reason behind the feature selection algorithms is to automatically select a set of features that are most relevant to the problem. The purpose of the feature selection is doubled: We want to improve model's performance and reduce model error by removing unnecessary features or noise. The size of acoustic and glottal features is huge, hence to avoid the case of over fitting and also, to avoid the curse of dimensionality we need to apply feature selection. Curse of dimensionality says that the complexity of a model grows exponentially with respect to the number of features. Thus, we just want a few numbers of features that is sufficient to generalize a model. For this we can use either use Sequential Forward Feature selection or Backward Feature selection method.

In this work [R1] [R2], we have used SFFS to select the essential features that will serve as an input to the classifier (Data-driven model).

- **Sequential Forward Feature selection(SFFS)**

  The sequential feature selection algorithms are a family of greedy search algorithms used to reduce the initial d-dimensional feature space in a k-dimensional feature subspace where k <d.

  Here we first select the best feature 'b' based on some scoring, max gain or min loss. Then, we are left with (d-1) features out of the'd' features. So, we compute the combined score of b and f_i for all i = 1 to d and i not equal to index of b, and then select the best score, hence obtaining the best two features. We keep doing this till a threshold value is reached. The pseudocode for the above explanation is given below:

```
def sequential_forward_feature_selection(actual_feature_list, d):
    selected_feature_list = [ ]
    b=argmax(scores(actual_feature_list))
    selected_feature_list.append(b)
    while threshold is not reached,do
        for i in range(d) and if i not in selected_feature_list, compute scores
       selected_feature_list.append(argmax(scores(selected_feature_list,f_i)))
    return selected_feature_list
```

6. Classification

The idea of classification is quite similar to the general ways of performing binary classification as shown below in fig. 2 and fig 3.

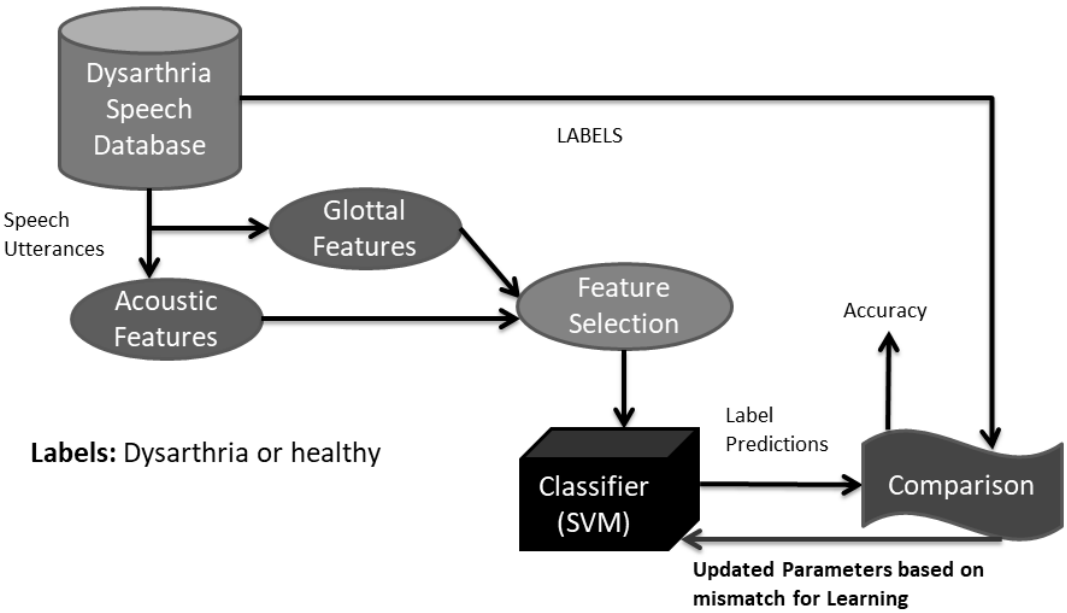

Fig. 2: Classification process for un-coded speech utterances.

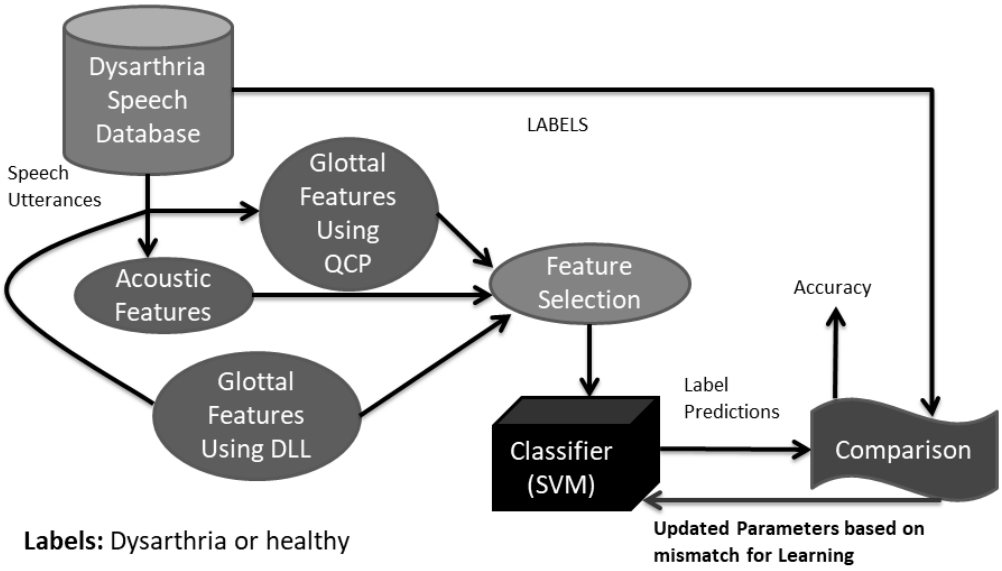

Fig. 3: Classification process for coded telephonic speech utterances.

Fig. 2 shows the classification approach [R1], where acoustic and QCP based glottal features are obtained from the speech utterance in the dysarthria speech database. Those extracted features are send to the SFFS based feature selector. Its output is then send to the SVM classifier and hence training is performed. The validation method used during training is K-Fold cross validation. Similar is the case for Fig. 3 which shows the classification approach [R2], for coded telephone speech. Here due to distortions in signals QCP based glottal features may not be useful, hence DNN-GIFs are used.

- **Support Vector Machines:**
  A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model set of labeled training data for each category, they're able to categorize new text. A support vector machine takes the data points and outputs the hyperplane (which in two dimensions it's simply a line) that best separates the tags. This line is the decision boundary: anything that falls to one side of it we will classify as Class 1, and anything that falls to the other as Class 2.

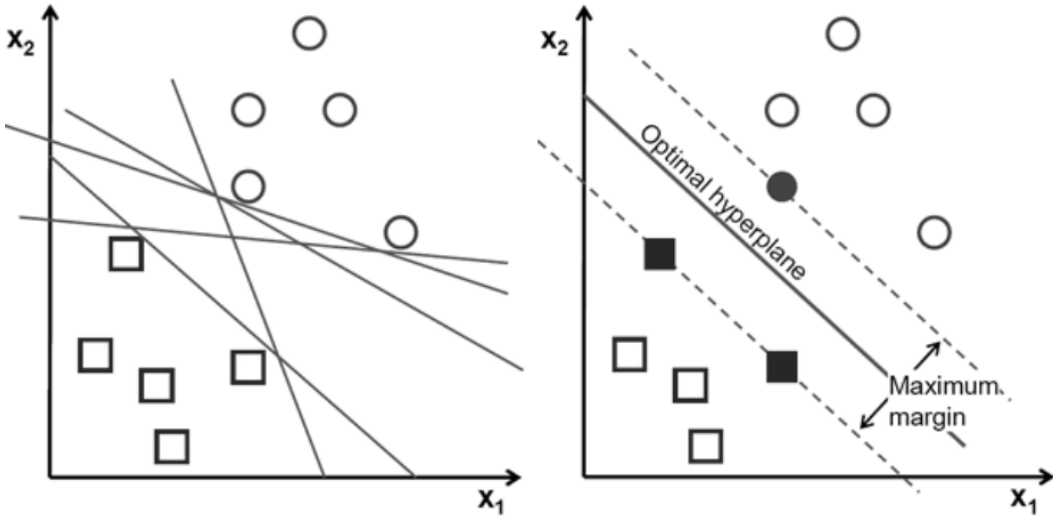

Fig 4: The optimal hyper plane in SVM.

For SVM, the best hyper plane is the one that maximizes the margins from both tags.

7. Observations and Conclusions

The tables below (table 3 and table 4) shows the importance of glottal features obtained from QCP for un-coded raw speech and the importance of DNN-GIF for the coded telephone signals (CTS) in improving the classification of dysarthria speech.

Feature set (Non-words)	Without Feature selection (%)	With Feature selection (%)
OpenSMILE-1	60.11	84.46
OpenSMILE-2	70.07	89.23
Glottal	69.34	78.16
OpenSMILE-1 + glottal	69.53	88.41
OpenSMILE-2 + glottal	67.75	93.52
Feature set (Words)	Without Feature selection (%)	With Feature selection (%)
OpenSMILE-1	78.84	88.39
OpenSMILE-2	80.36	93.39
Glottal	68.30	72.77
OpenSMILE-1 + glottal	77.14	92.77
OpenSMILE-2 + glottal	82.32	94.29
Feature set (Sentences)	Without Feature selection (%)	With Feature selection (%)
OpenSMILE-1	69.39	87.08
OpenSMILE-2	76.77	90.87
Glottal	61.08	71.86
OpenSMILE-1 + glottal	64.31	87.56
OpenSMILE-2 + glottal	74.63	91.38

Table 3: Classification accuracies for combination of different features with and without feature selection for un-coded raw speech data [R1].

Hence, we can conclude that the idea of extracting Glottal Features for dysarthria speech classification was a success and they contributed quite well to the improvement of classification accuracy.

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

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

Table 4: Classification accuracies for combination of different features with and without feature selection for coded telephone speech (CTS) data [R2].

8. Abbreviations

[R<n>] – refer to the nth article in the reference section.

9. References

1. N.P. Narendra et al, "Dysarthric speech classification using glottal features computed from non-words, words and sentences" - *Proceedings of Interspeech, 10.21437/Interspeech.2018-1059.*

2. N.P. Narendra et al, "Dysarthric speech classification from coded telephone speech using glottal features" - *Speech Communication 110 (2019) 47–55.*

3. Manu Airaksinen et al,"Quasi closed phase analysis for glottal inverse filtering*" – Research Gate Conference Paper, August 2013.*