Mittapalle, Kiran; Helkkula, Pyry; Keerthana, Y. Madhu; Kaitue, Kasimir; Minkkinen, Mikko; Tolppanen, Heli; Nieminen, Tuomo; Alku, Paavo

The Automatic Detection of Heart Failure Using Speech Signals

Published in:
Computer Speech and Language

DOI:
10.1016/j.csl.2021.101205

Published: 01/09/2021

Document Version
Publisher's PDF, also known as Version of record

Computer Speech & Language 69 (2021) 101205

Contents lists available at ScienceDirect

Computer Speech & Language

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

The automatic detection of heart failure using speech signals

M. Kiran Reddy*[,a], Pyry Helkkula[b], Y. Madhu Keerthana[c], Kasimir Kaitue[a], Mikko Minkkinen[d], Heli Tolppanen[d], Tuomo Nieminen[d], Paavo Alku[a]

[a] Department of Signal Processing and Acoustics, Aalto University, Aalto 00076, Finland
[b] Institute of Molecular Medicine, University of Helsinki, Helsinki 00290, Finland
[c] Advanced Technology Development Centre, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India
[d] Heart and Lung Center, Helsinki University Hospital, Helsinki 00290, Finland

ARTICLE INFO

Article History:
Received 20 August 2020
Revised 18 February 2021
Accepted 23 February 2021
Available online 27 February 2021

Keywords:
Heart failure
Mel-frequency cepstral coefficients
Glottal source parameters
Support vector machines
Extra tree
AdaBoost
Neural networks

ABSTRACT

Heart failure (HF) is a major global health concern and is increasing in prevalence. It affects the larynx and breathing – thereby the quality of speech. In this article, we propose an approach for the automatic detection of people with HF using the speech signal. The proposed method explores mel-frequency cepstral coefficient (MFCC) features, glottal features, and their combination to distinguish HF from healthy speech. The glottal features were extracted from the voice source signal estimated using glottal inverse filtering. Four machine learning algorithms, namely, support vector machine, Extra Tree, AdaBoost, and feed-forward neural network (FFNN), were trained separately for individual features and their combination. It was observed that the MFCC features yielded higher classification accuracies compared to glottal features. Furthermore, the complementary nature of glottal features was investigated by combining these features with the MFCC features. Our results show that the FFNN classifier trained using a reduced set of glottal + MFCC features achieved the best overall performance in both speaker-dependent and speaker-independent scenarios.

1. Introduction

Pathophysiologically, heart failure (HF) is characterized as the inability of the heart muscle to pump blood as well as it should (Coronel, 2001). HF is estimated to affect more than 37 million people worldwide (Ziaeian and Fonarow, 2016). Patients with HF experience various symptoms, such as edema, fatigue, dyspnea, rapid or irregular heartbeat, chest pain, and nausea (Ponikowski, 2016). These symptoms limit patients daily physical and social activities and result in poor quality of life (Zambroski, 2005; Rector, 2006). Although HF is a serious condition that progressively deteriorates over time, the majority of cases can greatly benefit from treatment at the right time.

The presence of edema, swelling caused by fluid retention in body tissues, is a distinctive feature of HF (Murton, 2017). HF-related edema causes an increase in weight as decompensation approaches (Joseph, 2009). Regular weighing is an easy and non-invasive way for patients to detect a weight increase at home (Joseph, 2009). However, edema-related weight gains occur relatively late in the disease progression and may not be detected in time to prevent decompensation (Joseph, 2009). Therefore, HF specialists emphasize the need for alternative non-invasive technologies that can identify edema earlier than weight gain (Gheorghiade, 2005). Speech is an emerging non-invasive biomarker that has been associated with several pathologies including dysphasia, obstructive sleep apnea, and other neurological and laryngeal disorders (Kiran Reddy et al., 2020;

*Corresponding author.
E-mail address: kiran.r.mittapalle@aalto.fi (M. Kiran Reddy).

https://doi.org/10.1016/j.csl.2021.101205

Goldshtein, 2011; Orozco-Arroyave, 2015). Two recent studies have investigated the relationship between speech and heart disease (Murton, 2017; Maor, 2016). In Murton (2017), the authors analyzed how several measures of voice quality, such as the cepstral peak prominence (CPP) parameter, correlate with improvements in HF symptoms during a decompensation treatment. Logistic regression analysis of mel-frequency cepstral coefficients (MFCCs) was used to study the association between coronary artery disease and voice characteristics in Maor (2016). These studies showed for the first time that non-invasive voice signal characteristics are associated with heart diseases. Furthermore, it has been observed that the severity of HF-related edema required to measurably change the voice is small compared to the severity needed to increase body weight (Murton, 2017). Hence, speech-based approaches have a great potential to be used for non-invasive detection of HF at an early stage.

HF-related edema in the vocal folds and lungs is hypothesized to affect phonation and speech respiration (Murton, 2017). Therefore, the acoustic features extracted from two of the main parts of the speech production process, glottal excitation and vocal tract, may contain useful information for identifying HF. To the best of our knowledge, there are no previous studies on the automatic detection of HF from speech signals. The current study focuses on investigating the effectiveness of MFCC features, glottal features, and their combination in distinguishing the speech of HF patients from the speech of healthy controls. The glottal features consist of time- and frequency-domain glottal parameters, which characterize different aspects of the airflow excitation waveform of voiced speech, the glottal waveform (Alku, 2002). While MFCCs mainly capture the vocal tract information of speech, the glottal features represent the source information of the speech production mechanism. The features extracted from every speech utterance, as well as the corresponding binary label indicating the presence of HF (i.e., *HF* vs. *healthy*), were used to train support vector machine (SVM), Extra Tree (ET), AdaBoost, and feed-forward neural network (FFNN) classifiers. The performance of the classifiers was compared in speaker-dependent and speaker-independent scenarios. The speaker-independent classification was performed on training data by using a leave-five-speakers-out cross validation strategy. The experimental results show that FFNN provides the best classification performance when trained using the optimal subset of features obtained via feature selection.

The rest of the paper is structured as follows. Section 2 describes the proposed method, the collected speech database, the acoustical features, and the classifiers used in the current study for classifying healthy from HF speakers. In Section 3, statistical analyses of the glottal parameters between healthy speakers from HF speakers are reported. Further, this section also demonstrates the acoustical changes in the speech signal and glottal excitation, caused by HF. The results of the classification experiments are described in Section 4. Section 5 summarizes the present work.

2. The proposed method

Fig. 1 depicts the steps in the proposed system for the automatic detection of HF from a speech signal. In the training phase, the features extracted from speech signals and the corresponding *HF*/*healthy* labels are used to train four classifiers, namely, SVM, ET, AdaBoost, and FFNN classifiers. Two types of feature sets are extracted from every speech utterance present in the speech corpus (described in Section 2.1). The first set consists of glottal parameters extracted from glottal flow waveforms. The glottal waveforms are estimated from speech signals by using a recently proposed automatic glottal inverse filtering (GIF) algorithm, the quasi-closed phase (QCP) method (Airaksinen, 2013). The QCP method is chosen for GIF because it was shown to perform better than the four most popular GIF algorithms (Airaksinen, 2013). From the obtained glottal flow waveforms, 12 time- and frequency-domain glottal parameters are extracted using the APARAT toolbox (described in Section 2.2). The second set consists of MFCC features and their statistics (described in Section 2.3). The terms *feature set* and *features* are used interchangeably in this paper.

The SVM, ET, AdaBoost, and FFNN classifiers (discussed in Section 2.4) are trained using the features extracted from speech utterances as input and the corresponding labels as output. Each classifier is trained separately using individual and combined (MFCC and glottal) feature sets. During testing, the trained classifiers can be used to detect the presence of HF from the speech signals. The same set of speech features that were used during training, are extracted from the test speech utterance. The extracted features are given as input to the classifiers, and the classifier predicts the labels (*HF*/*healthy*).

2.1. The HF database

As part of the current study, a new speech database was recorded by the authors. The database includes recordings in Finnish (text readings and spontaneous speech) by 25 healthy speakers (7 female and 18 male) and 20 HF patients (6 female and 14 male). The ages of the healthy speakers were between 54 and 83 (mean: 60) and the ages of the HF patients were between 36 and 81 (mean: 67). The patients were hospitalized for HF of any etiology, regardless of the left ventricular ejection fraction. The speech data, sampled at 44.1 kHz, were recorded in doctor's practice rooms using a condenser microphone (DPA 4065-BL). Each speaker read the same Finnish text of 91 words three times (the text-reading task) and produced one spontaneous speech. For the experiments of the current study, we have only considered the middle recitation of the text-reading task. This implies that the language and linguistic content of the speech is the same for all the speakers. A linear phase FIR filter (cut-off frequency: 60 Hz) was used to remove the low-frequency noise picked up by the recording microphone. The duration of each speech file was $\sim$1 min. The wave files were chopped into non-overlapping segments of 4 s each. Hence, a total of 628 WAV files (segments) were available for analysis.

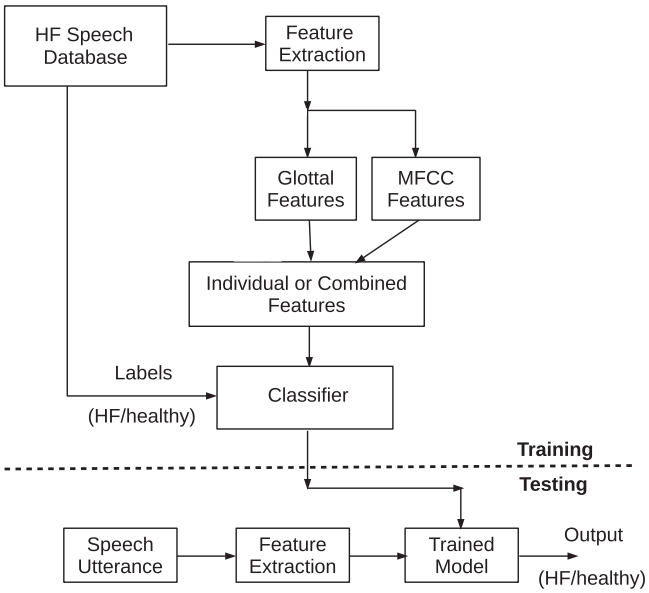

Fig. 1. The proposed method for HF detection from speech signals.

2.2. Extraction of the glottal feature set

The glottal flow waveform estimated by using the QCP was parameterized with a glottal parameter set consisting of 12 known time- and frequency-domain parameters (Childers and Lee, 1991; Alku, 2002). These parameters characterize various aspects of the glottal flow waveform, and were estimated using the APARAT toolbox (Airas, 2005). The glottal parameters are listed in Table 1. The glottal parameters were computed in 30 ms frames. While the harmonic richness factor (HRF) and the difference between the first two glottal harmonics (H1H2) were computed pitch asynchronously once per frame, the remaining parameters were computed pitch synchronously once per glottal cycle and then averaged over the frame. All of the nine time-domain parameters and the parabolic spectral parameter (PSP) were expressed using a linear scale, while H1H2 and HRF were expressed using the dB scale. The glottal parameters computed from all voiced frames of the input speech signal finally form the glottal parameter vector of the utterance. Voicing detection was computed using a simple method based on the frame's log energy. The following eight statistical measures were computed from the glottal parameter vector, as well as from its first derivative: minimum, maximum, mean, median, standard deviation, range, kurtosis, and skewness. This results in $(12 + 12) \times 8 = 192$ parameters, referred to in this work as the glottal feature set (GFS).

Table 1
Time- and frequency-domain glottal parameters. For more details, see Childers and Lee (1991) and Alku (2002).

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

2.3. Extraction of the MFCC feature set

In this study, 13 MFCCs were extracted every 10 ms. The MFCCs computed from all the voiced frames of the input speech signal form the MFCC parameter vector of the utterance. Eight statistical measures were computed from the MFCC parameter vector: minimum, maximum, mean, median, standard deviation, range, kurtosis, and skewness. This results in $13 \times 8 = 104$ parameters representing the MFCC feature set (MFS).

2.4. Classifiers

2.4.1. SVM

One of the most prevalent supervised learning models is SVM (Cortes, 2010). This algorithm aims to find the optimal hyperplane, based on labeled training data, which can be used to classify the test data. Although deep learning has become a prevalent approach in many classification studies in speech technology (such as recurrent neural networks (RNNs) for emotion recognition (Mirsamadi, 2017), convolutional neural networks (CNNs) for speaker identification (An et al., 2019), generative adversarial nets (GANs) for language identification (Shen, 2017), and convolutional LSTM deep neural networks (CLDNNs) for spoof detection (Dinkel et al., 2018)), SVM still constitutes a highly justified classifier technology in detecting speech from patients with diseases such as dysphasia (Kiran Reddy et al., 2020), obstructive sleep apnea (Goldshtein, 2011), predementia, and Alzheimer's (König, 2015). This is because speech databases recorded from patient populations usually contain little data, and the amount of training data is typically not sufficient for training data-hungry deep learning networks. In this work, the Scikit-learn Python library (Pedregosa, 2011) was used to implement the non-linear SVM algorithm with a radial basis function (RBF) kernel. The RBF kernel was chosen by first comparing its performance with two other kernels (linear and polynomial) in pilot experiments and by observing that the RBF kernel yielded the best accuracy among the three kernels compared. The kernel equation is given by

$$K(x,y) = exp(-\gamma \parallel x - y \parallel^2), \quad \gamma > 0, \tag{1}$$

where x and y are training samples and labels, respectively, and γ is the kernel parameter. In addition to γ, regularization was used in the SVM with a regularization parameter C.

2.4.2. ET

The ET classifier (Geurts, 2006), a variant of Random Forest, is a type of ensemble learning technique that aggregates the results of multiple de-correlated decision trees to output its classification result. Unlike the Random Forest method, the ET classifier splits nodes by choosing cut-points fully at random, and it also uses the whole learning sample (rather than a bootstrap replica) to grow the trees. The ET algorithm works by creating a large number of unpruned decision trees from the training data according to the classical top-down procedure (Geurts, 2006). Predictions are made by using majority voting from decision trees. The main advantage of the ET algorithm is that it gives results that are comparable to those obtained by Random Forest but this is achieved computationally more effectively (Geurts, 2006). The ET classifier was implemented in the current study using the Scikit-learn Python library (Pedregosa, 2011).

2.4.3. AdaBoost

Adaptive boosting (AdaBoost) (Freund, 1999) is an iterative ensemble boosting approach that constructs a strong classifier sequentially by combining multiple weak classifiers. The set of weak classifiers is built iteratively from the training data. At each iteration, a higher weight is assigned to wrongly classified samples so that these observations get a high probability for classification in the succeeding iteration. Weights are also computed for the classifiers according to the classification accuracy. A more accurate classifier will get a higher weight. During testing, the predicted class of an input sample is computed as the weighted mean prediction of the classifiers in the ensemble. The AdaBoost classifier was also implemented using the Scikit-learn Python library (Pedregosa, 2011).

2.4.4. FFNN

Another classifier explored in the current study is FFNN. This is a supervised learning algorithm that learns a non-linear function $f(\cdot) : R^m \rightarrow R^n$ from the training data, where m and n are the dimensions of input and output, respectively. Deeper networks require large amounts of data for proper training, which is generally not available for speech data recorded from patients (Kiran Reddy et al., 2020). On the other hand, a single hidden layer FFNN consisting of an appropriate number of hidden units may give the desired performance even with a small dataset (Kiran Reddy et al., 2020). Hence, in this work, an FFNN with a single hidden layer is used. The Keras framework Keras was used for constructing the FFNN. The optimal number of hidden neurons was derived using a grid search algorithm (described in Section 4.1). A rectified linear unit (ReLU) activation function was used for the hidden layer, and the output layer used a sigmoid activation function. The maximum number of epochs for training was set to 25. The learning rate and the mini-batch size were set to 0.001 and 32, respectively. The squared hinge loss was selected as the loss function. The weights of the network were initialized randomly, and they were optimized using the RMSprop algorithm.

3. Analysis of the glottal parameters and Log-Mel spectrograms between speakers with HF and their healthy controls

This section describes statistical tests that were computed from the glottal parameters (described in Table 1) in order to study whether they show significant differences between the two speaker groups (healthy speakers vs. speakers with HF). It should be noted (see Section 2.1) that the glottal parameters were computed from Finnish speakers reading the same text, that is, the language and linguistic content of the speech data are the same for all speakers in both groups. In addition, the two groups are similar in age, and none of the speakers had any known disorders (e.g., a common cold) that could have affected their speech production at the time of the recordings, except for HF in the case of the HF group. Therefore, it can be assumed that if the statistical tests showed a significant difference between the two speaker groups, the difference is expected to be due to HF.

Statistical tests were carried out with one-way repeated measures analyses of variance (ANOVAs) using each of the glottal parameters as a dependent variable and the speaker group (healthy speaker vs. HF patient) as an independent variable. The glottal parameters computed from all the voiced frames of an utterance were first averaged to obtain the utterance-level parameters for which the ANOVA tests were computed. The results of the ANOVA tests are given in Table 2. From the table, it can be observed that, except for PSP, all the glottal parameters showed statistically significant ($p < 0.01$) differences between the healthy speakers and the speakers with HF. In addition to the ANOVA test, the statistical distributions of the glottal parameters between the two speaker groups are described using box plots for four selected glottal parameters in Fig. 2. This figure shows the distributions of three time-domain glottal parameters (the open quotient, calculated from the primary glottal opening (OQ1), the closing quotient (ClQ), and the speed quotient, calculated from the secondary glottal opening (SQ2)) and one frequency-domain glottal parameter (H1H2). The box plots demonstrate that the glottal parameters discriminate HF patients from their healthy controls. In addition, these box plots enable describing the general differences in the glottal flow pulse between the HF patients and their healthy controls as follows. The three time-domain parameters shown in Fig. 2 demonstrate that the mean values of OQ1 and ClQ are larger for the speakers with HF compared to their healthy controls and that the mean value of SQ2 is closer to 1.0 for the speakers with HF compared to their healthy controls. As described in previous studies on the parameterization of human speech production (Holmberg et al., 1988; Alku, 2011), an increase in OQ1 and ClQ combined with the value of SQ2 moving close to 1.0 implies that the glottal flow pulse changes to a more rounded, soft shape. In the frequency domain, this corresponds to an increase in the spectral tilt of the glottal flow, an issue which is also demonstrated by Fig. 2 (i.e., the mean value of H1H2 is lower for the speakers with HF compared to their healthy controls). In summary, the ANOVA test results, along with the parameters' box plots, suggest that the speakers with HF tend to produce their speech by using a softer, low-frequency rich glottal flow pulse compared to their healthy controls.

Fig. 3 shows examples of glottal source waveforms estimated using the QCP from speech signals produced by a healthy speaker and by a speaker with HF. The figure demonstrates that the glottal flow of the healthy speaker shows a clear closed phase and a short open phase. However, for the speaker with HF, the closed phase is practically absent from the glottal flow and the shape of the glottal pulse is more rounded.

Finally, in order to demonstrate differences between speech signals produced by speakers with HF and their healthy controls, Fig. 4 shows time-domain waveforms and log-mel spectrograms over a time span that covers one sentence. A spectrogram, which is similar to the wavelet scalogram (Guariglia, 2017; Keerthana et al., 2019; Zheng et al., 2019; Mallat, 1989), is a time-frequency representation of the speech signal. In this example, a healthy male speaker (left panels) and a male speaker with HF (right panels) uttered the same written sentence. It can be seen that the spectrogram of the healthy speaker shows a clear harmonic structure, especially at low frequencies. In contrast, the spectrogram of the speaker with HF shows a blurred harmonic structure containing more noise-like components. Hence, the log-mel spectrogram contains information that enables discriminating

Table 2
One-way analysis of variance results for the glottal parameters using the speaker group (healthy speaker vs. HF patient) as an independent variable. Statistically significant values are shown in bold.

Parameter	*F*-value	*p*-value
NAQ	**9.672**	**0.001956**
AQ	**13.975**	**0.000202**
OQ1	**143.5137**	**0.000000**
OQ2	**8.703**	**0.003296**
QOQ	**97.025**	**0.000000**
OQa	**44.215**	**0.000000**
ClQ	**82.987**	**0.000000**
SQ1	**84.498**	**0.000000**
SQ2	**88.796**	**0.000000**
HRF	**88.796**	**0.000000**
PSP	3.484	0.076386
H1H2	**86.199**	**0.000000**

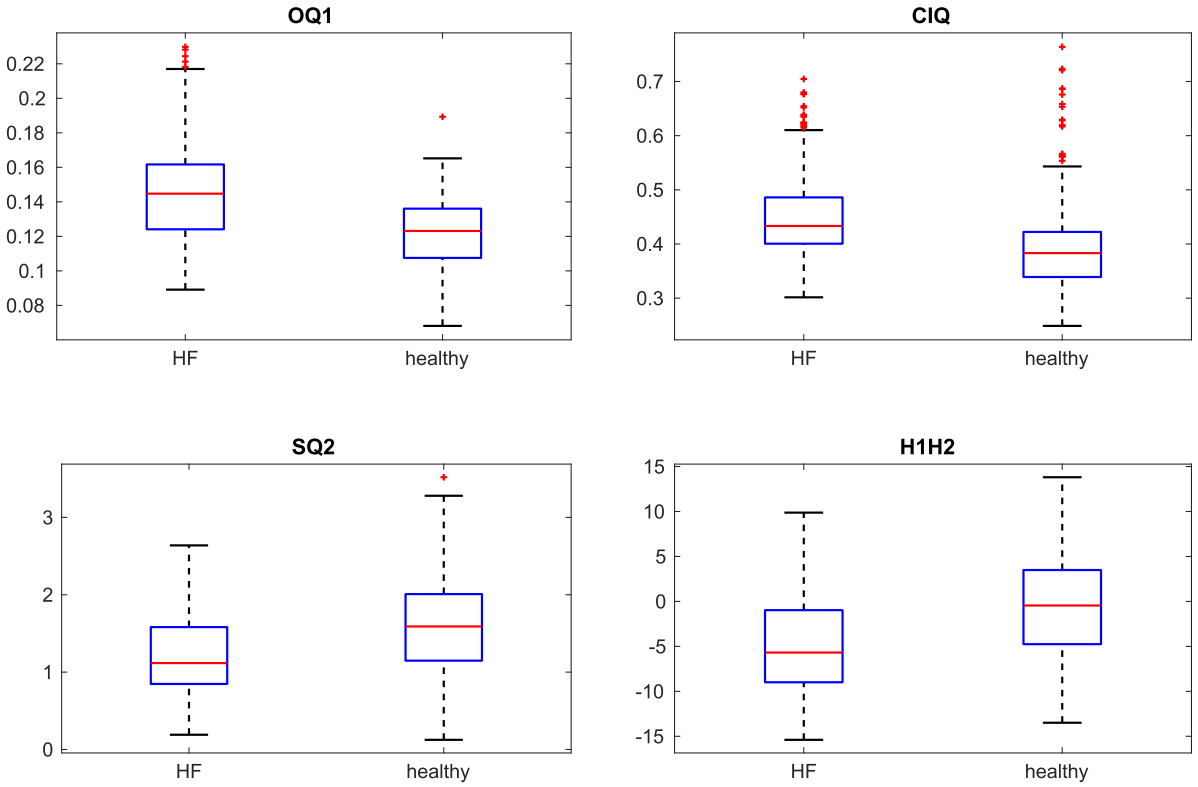

Fig. 2. The box plots of four glottal parameters computed using the QCP method and the speech of HF patients and their healthy controls. The central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles respectively. The whiskers on either side cover all points that are within 1.5 times the interquartile range, and points beyond these whiskers are plotted as outliers using the "+" symbol.

speakers with HF from their healthy controls. In the current study, the MFCCs were used to express the log-mel spectrogram. The use of MFCCs can be justified by the observations above but also by the results of a previous study (Maor, 2016), where MFCCs were used in the detection of coronary artery disease from speech signals.

4. Classification experiments and results

The main focus of the current study is on investigating whether the different machine learning classifiers and acoustical features described in Section 2 are capable of distinguishing the speech of HF patients from the speech of healthy controls. In order to study this, classification experiments were arranged, and they will be described in Section 4.1. The results of the classification experiments are reported in Section 4.2.

4.1. The experimental setup

A grid search algorithm was used to find the optimal parameter values for SVM. For this, speech utterances from three HF speakers and three healthy speakers were used as validation data, and the speech utterances from the remaining speakers were used as training data. To perform the grid search, γ and C were varied from 10^{-3} to 10^{3} in multiples of 10. The combination (γ, C), which achieved the best validation performance, was chosen. The same procedure was followed to determine the optimal number of hidden neurons in FFNN by varying the number of neurons from 128 to 1024 in steps of 128. It was observed that SVM achieved the best validation performance using the parameter combination $\gamma = 0.01$ and $C = 10$. FFNN performed best when the number of hidden units was set to 512. The two hyper-parameters of AdaBoost (the number of estimators and the learning rate) were also set using a grid search. The number of estimators was varied from 50 to 500 in steps of 25, and the learning rate was varied from 10^{-3} to 10^{1} in multiples of 10. The ET algorithm has several hyper-parameters, but tuning too many hyper-parameters could also make Extra Trees too similar to Random Forest. Therefore, the grid search was used to find optimal values for the three most important hyper-parameters: the number of trees in the forest, the minimum number of samples required to split an internal node, and the minimum number of samples required to be at a leaf node. The number of trees was varied from 50 to 500 in steps of 25. The other two parameters were varied from 2 to 20 in steps of 2. The combination of the three hyper-parameters, which achieved the best validation performance, was finally chosen. The individual features were standardized by applying zero mean and unit variance normalization before applying the machine learning models.

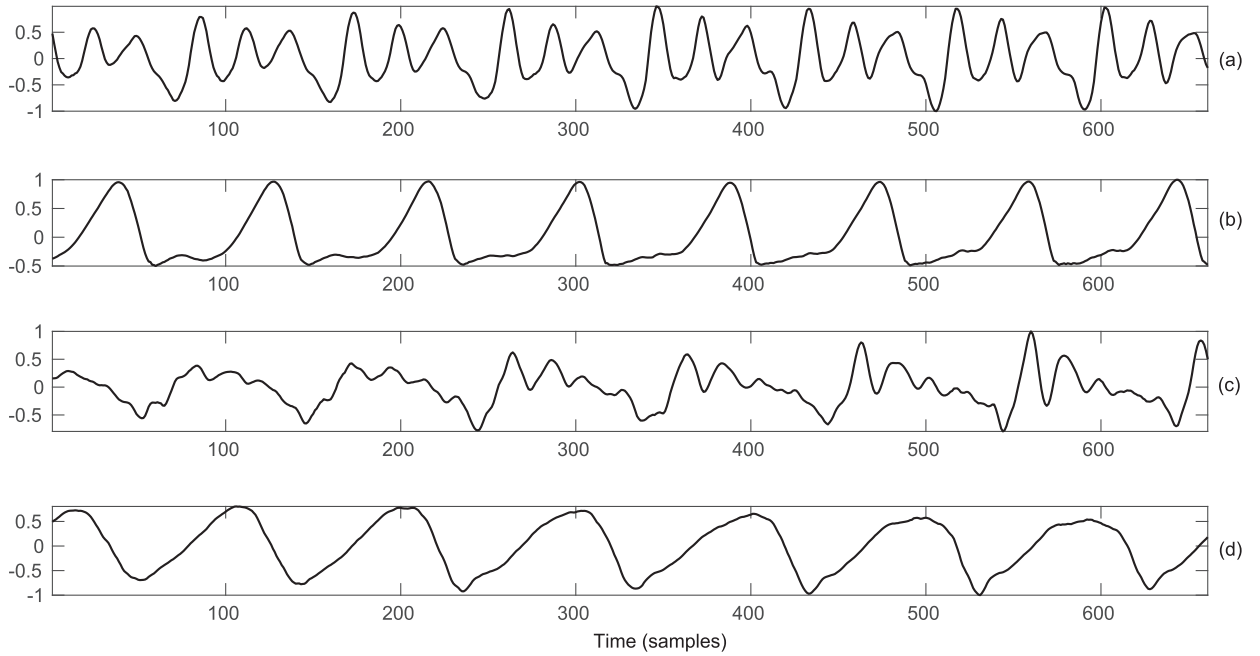

Fig. 3. An illustration of glottal flow waveforms estimated using the QCP from speech signals produced by a healthy speaker and by a speaker with HF. The signals shown in panels are: (a) the acoustic speech signal produced by the healthy speaker, (b) the corresponding glottal flow waveform, (c) the acoustic speech signal produced by the speaker with HF, and (d) the corresponding glottal flow waveform. The speech signals were extracted from parts of the vowel [a].

There were 45 speakers altogether (25 healthy speakers and 20 HF patients) in the database. In the speaker-dependent scheme, 70% of the wave files (randomly chosen) were used for training, and the remaining 30% were used for testing. The experiment was repeated ten times, each time building different training and testing sets. The final classification accuracy was obtained as the average of accuracy obtained at each iteration. In the speaker-independent scheme, a leave-five-speakers-out cross validation strategy was used to determine the classification accuracy of the training data. In this strategy, five speakers were used at every fold for validation, and all other speakers were used for training. The cross-validation process was then

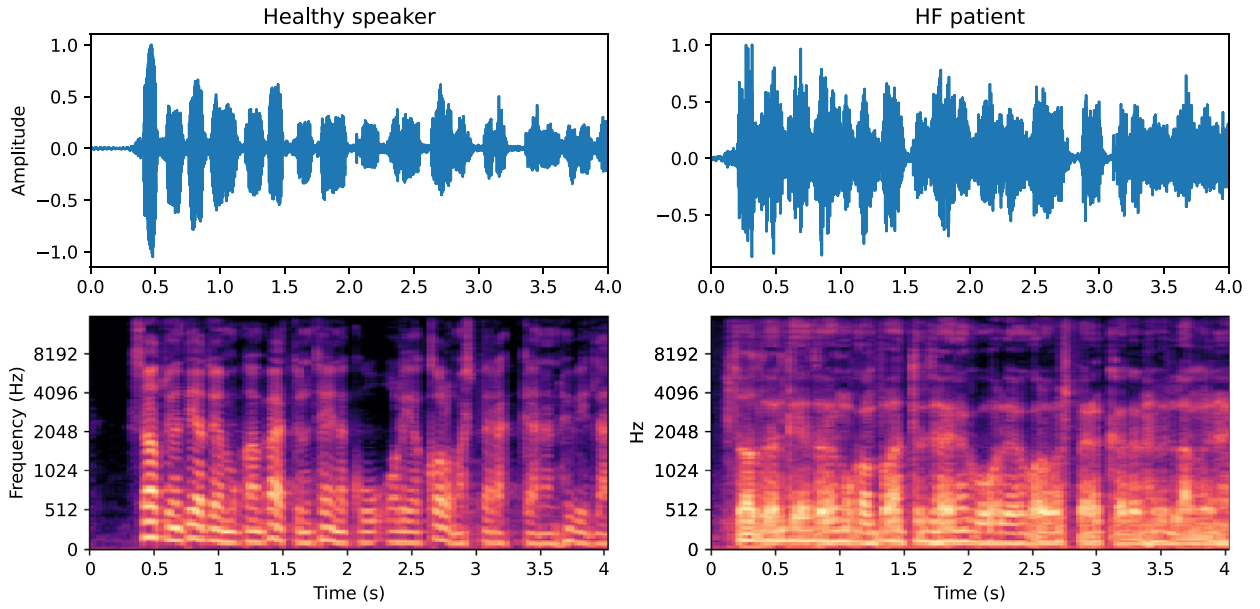

Fig. 4. An illustration of differences in speech signals between a healthy speaker (left panels) and a speaker with HF (right panels), both producing the same sentence. The top panels show the time-domain speech signals and the bottom panels show the corresponding log-mel spectrograms.

Table 3
Average classification accuracies (%) obtained using various individual feature sets in speaker-dependent mode. The feature vector dimension is given in parentheses.

Classifier	MFS (104)	GFS (192)	MFS + GFS (296)	MFS + GFS with feature selection (85)
SVM	91.34	77.89	90.23	94.47
ET	88.58	76.71	85.32	91.84
AdaBoost	84.77	75.36	86.23	88.41
FFNN	90.56	75.91	89.66	**95.02**

repeated with each set of five speakers used exactly once for validation. The classification accuracies obtained at all folds were averaged to obtain the final accuracy. The classification accuracy was computed as the ratio of the number of correctly classified speech utterances to the total number of speech utterances.

4.2. Results

The average classification accuracies obtained using the individual feature sets for the speaker-dependent case are shown in Table 3. From the table, it can be observed that the classification accuracies obtained with the MFS set, which is based on MFCCs, are better than the ones obtained using the GFS set, which uses glottal features, for any classifier. Moreover, it can be seen that when MFS is used in different classifiers, SVM achieves the highest classification accuracy. Given the results reported in Section 3, which indicate that the glottal features are significantly affected by HF, and also given the fact that MFCCs were originally developed to parameterize the vocal tract characteristics of speech signals (Davis and Mermelstein, 1980), the better performance of MFCCs compared to the glottal features in the classification experiments might at first seem surprising. Similar results have, however, also been recently observed in other SVM-based biomarking studies [8]. The potential explanations for the better performance of MFCCs compared to the glottal features are as follows. First, despite the fact that MFCCs focus on the parameterization of the vocal tract, they also carry information about the glottal source, such as irregular movements and incomplete closure of the vocal folds in the presence of pathologies (Godino-Llorente et al., 2006). Second, it is possible that HF affects not only the glottal flow (as reported in Section 3) but also the vocal tract by causing incomplete closure of the vocal folds, which induces changes to formant bandwidths and the spectral tilt of the tract when compared with the same vocal tract geometry with a fully closed glottis (Barney, 2007). Therefore, MFCCs are also able to parameterize the vocal tract changes caused by HF while these changes do not affect the glottal features. These two reasons explain why the cepstral feature vector given by the MFCC computation turned out to be more effective in the detection task compared to the corresponding glottal feature vector.

An improvement in the classification accuracy was achieved by combining the glottal features with MFS. This shows that the glottal features and MFCCs carry complementary HF-related information. Although the accuracy obtained with MFS + GFS is higher than with the glottal features, it was still less compared with MFS. This may be due to the presence of irrelevant and redundant features in the combined feature set. Therefore, we applied feature selection on the combined feature set in order to choose an appropriate subset of features to improve the accuracy. The feature selection was conducted using an ET classifier. To perform feature selection, the normalized total reduction in the Gini impurity (Sandri and Zuccolotto, 2008) was computed for each feature, during the construction of the ET forest. This value is called the Gini importance of the corresponding feature. Only those features whose Gini importance is greater than the mean Gini importance of all features are retained, and the remaining features are eliminated. The bar plot in Fig. 5 demonstrates the Gini importance for the various features and how the selection procedure selects features from the MFCC and glottal source feature sets. From the last column of Table 3, it can be seen that the classification accuracy further improved after the feature selection. Furthermore, the feature selection has drastically reduced the size of the combined feature set. This enhances the computational efficiency as well as the generalization capabilities. The best classification accuracy (95.02%) was achieved when FFNN was trained using the reduced set of the MFS + GFS features.

Table 4 shows the average classification accuracies in the speaker-independent scenario. The results show a trend similar to the ones in Table 3. Although the classification accuracies are low compared to those achieved in the speaker-dependent mode, the models performed fairly well on unseen test data. Even in the speaker-independent case, FFNN trained using the reduced set of MFS + GFS achieved the best classification accuracy (81.51%).

The best performing FFNN classifier in Table 3 was further analyzed using a confusion matrix in a speaker-independent evaluation. This was done to make sure that the models were not biased towards a particular class. For these experiments, 15 HF and 15 healthy speakers selected randomly from the database were considered for training, and the remaining 15 speakers (five HF and ten healthy speakers) were used for testing. The data in the training and test set were normalized and shuffled randomly. The experiment was repeated 20 times, each time building different training and testing sets. There was no overlap in the data used during training and testing in any iteration. The final results are presented in a confusion matrix, where various measures are defined as in Saenz-Lechon (2006):

1. *The true positive rate (tp)* (which is also called sensitivity) is the ratio between the number of HF utterances correctly classified and the total number of HF utterances.
2. *The false negative rate (fn)* is the ratio between the number of HF utterances wrongly classified and the total number of HF utterances.

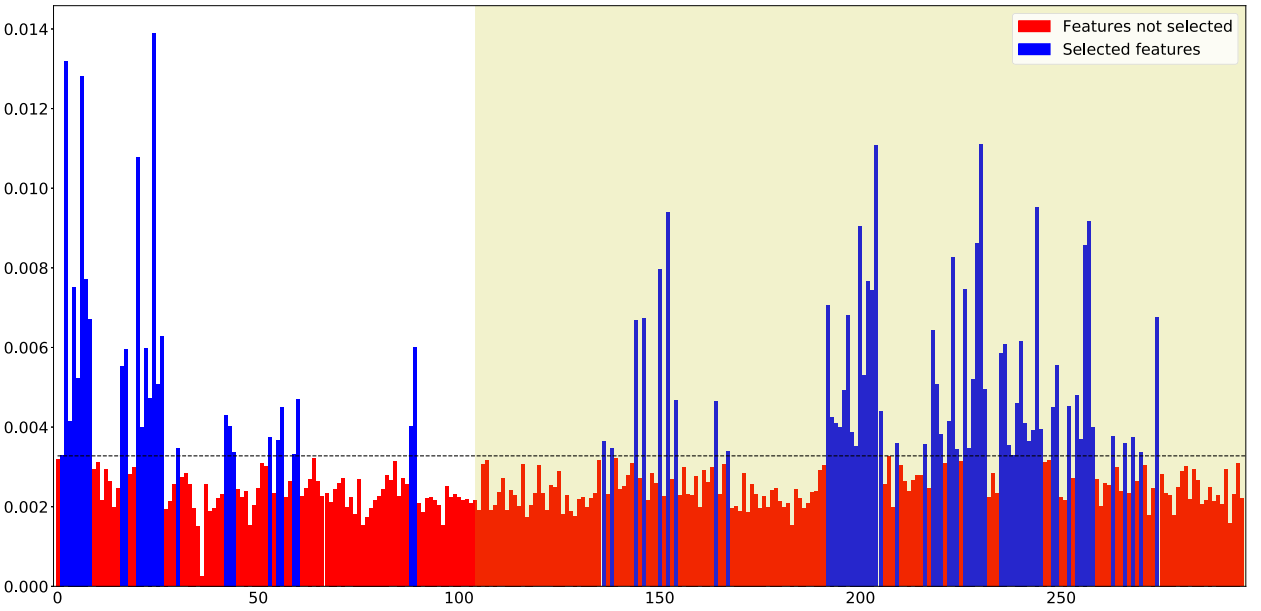

Fig. 5. Bar chart of the ET classifier's Gini importance score for each feature. The *x*-axis indicates the index of the individual feature in the combined feature vector. The unshaded and light-yellow shaded portions show the scores for MFCC and glottal source descriptors, respectively. The dotted line indicates the threshold for feature selection. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. *The true negative rate (tn)* (which is also called specificity) is the ratio between the number of healthy utterances correctly classified and the total number of healthy utterances.
4. *The false positive rate (fp)* is the ratio between the number of healthy utterances wrongly classified and the total number of healthy utterances.

Table 5 shows the confusion matrices with the mean and standard deviation values obtained by averaging the results for each individual iteration. From the table, it can be seen that both classes are being predicted equally well with the FFNN classifier. The overall results indicate that the FFNN trained using the reduced MFS + GFS set provides the best performance in the HF detection from speech signals.

In order to get a preliminary understanding of whether the proposed system can be used in real-life scenarios, the FFNN and SVM classifiers were further studied in the detection of HF from noisy speech. The classifiers were trained using clean speech and tested using noisy speech (i.e., the experiments were conducted under mismatched noise conditions). As discussed in Section 4.1, 70% of the data was used for training and the remaining 30% for testing. The test data was corrupted by additive noise in different signal-to-noise (SNR) conditions using two types of noise: (1) stationary office noise and (2) non-stationary traffic noise. Since the classifiers were shown to perform best with the reduced MFS + GFS feature set in the preceding experiments, this feature set was used here to train both classifiers. Fig. 6 shows the classification accuracies obtained with the SVM and FFNN classifiers at different SNR values and also for clean test speech. As mentioned in Section 2.1, the speech signals were recorded in doctor's practice rooms, and hence a small amount of ambient noise is already present in the clean speech. This ambient noise was removed using a linear phase FIR filter. It can be seen from the figure that the performance of the SVM and FFNN classifiers is better in the case of office noise compared to traffic noise. This is because office noise is stationary with predominantly low-frequency components. On the other hand, traffic noise is non-stationary, that is, it is time-varying in nature, and it also has more high-frequency energy. Compared to MFCCs, the glottal features are expected to contribute to a fair reduction in performance due to the inherent

Table 4
Average classification accuracies (%) obtained using various individual feature sets in speaker-independent mode. The feature vector dimension is given in parentheses.

Classifier	MFS (104)	GFS (192)	MFS + GFS (296)	MFS + GFS with feature selection (85)
SVM	76.26	65.59	75.52	77.15
ET	68.07	64.15	64.80	72.29
AdaBoost	76.54	68.85	73.04	79.23
FFNN	77.02	71.03	75.34	**81.51**

Table 5
Confusion matrix using MFS + GFS with feature selection. Average classification (%) (mean ± standard deviation).

Actual Labels	MFS + GFS	
	HF	Healthy
HF	79.47 ± 4.41 *(tp)*	20.53 ± 4.41 *(fn)*
Healthy	17.31 ± 4.16 *(fp)*	82.69 ± 4.16 *(tn)*

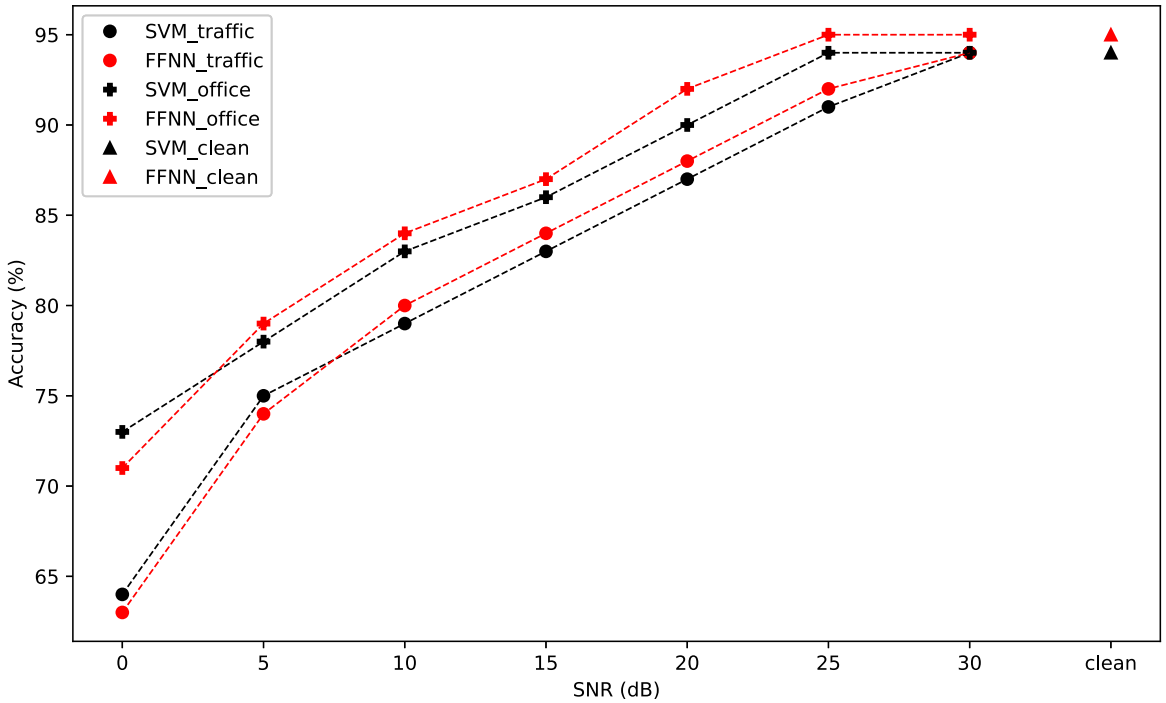

Fig. 6. The classification accuracy of the FFNN and SVM classifiers trained with clean speech and tested with noisy speech of different SNR values. Two types of additive noise (traffic and office) were used. As a reference, accuracy in the case of clean test speech is also shown.

sensitivity of GIF to artefacts (Alku, 2011). While there is a sharp decrease in accuracy for both classifiers at low SNRs, there is only a small decrease in accuracy when the noise condition changes from clean to mild and moderate (i.e., SNR conditions between 15 and 30 dB). This indicates that the proposed approach can potentially be used in realistic conditions where the input speech signal is not clean but mildly degraded by noise.

5. Summary and conclusion

HF is a serious and costly health problem, which remains one of the leading causes of death worldwide. Hence, there is a need for technology to biomark HF at an early stage and to detect even a slight decompensation. Speech offers a biomarker that is inherently cost-effective, comfortable, and non-invasive. Speech biomarkers have shown the potential to detect several disorders, but the speech-based detection has not been studied previously for HF. In this paper, a preliminary study is reported to elucidate on the potential of using speech in the biomarking of HF.

The proposed method utilizes the MFCC and glottal features extracted from speech to discriminate HF patients from their healthy controls. The MFCC and glottal feature sets are extracted for every speech utterance. The glottal features are extracted from the glottal flow waveforms estimated using the QCP inverse filtering method. The experimental results show that the MFCC features achieved a higher classification accuracy compared to the glottal features. The classification accuracy achieved with the combined (MFCC and glottal) feature set was comparable (although slightly inferior) when compared to the MFCC feature set. Most importantly, the results indicate that each classifier achieved the highest classification accuracy when trained using the reduced set of MFS + GFS. This confirms that the glottal and MFCC features carry discriminative information, which is essential

for the development of robust HF detection systems. Our experiments also showed that FFNN performs better than the SVM, ET, and AdaBoost classifiers when trained using the reduced of MFS + GFS features. The proposed method demonstrated the effectiveness of the MFCC and glottal features, in both speaker-dependent and speaker-independent speech-based HF detection. In the future, the proposed method can be extended to detect the stages and types of HF.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research has been funded by the Academy of Finland (project no. 330139)

References

Airaksinen, M., et al., 2013. Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (3), 596–607.
Airas, M., et al., 2005. A toolkit for voice inverse filtering and parametrisation. pp. 2145–2148.
Alku, P., et al., 2002. Normalized amplitude quotient for parameterization of the glottal flow. J. Acoust. Soc. Am. 112 (2), 701–710.
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 36, 623–650.Part 5
An, N.N., Thanh, N.Q., Liu, Y., 2019. Deep CNNs with self-attention for speaker identification. IEEE Access 7, 85327–85337.
Barney, A., et al., 2007. The effect of glottal opening on the acoustic response of the vocal tract. Acta Acust. United Acust. 93 (6), 1046–1056.
Childers, D.G., Lee, C.K., 1991. Vocal quality factors: analysis, synthesis, and perception. J. Acoust. Soc. Am. 90 (5), 2394–2410.
Coronel, R., et al., 2001. Defining heart failure. Cardiovasc. Res. 50 (3), 419–422.
Cortes, C., et al., 2010. Two-stage learning kernel algorithms. In: Proc. International Conference on Machine Learning, pp. 239–246.
Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. 28 (4), 357–366.
Dinkel, H., Qian, Y., Yu, K., 2018. Investigating raw wave deep neural networks for end-to-end speaker spoofing detection. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (11), 2002–2014.
Freund, Y., et al., 1999. A short introduction to boosting. J. Jpn. Soc. Artif.Intell. 14 (5), 771–780.
Geurts, P., et al., 2006. Extremely randomized trees. Mach. Learn. 63 (1), 3–42.
Gheorghiade, M., et al., 2005. Acute heart failure syndromes: current state and framework for future research. Circulation 112 (25), 3958–3968.
Godino-Llorente, J.I., Gomez-Vilda, P., Blanco-Velasco, M., 2006. Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters. IEEE Trans. Biomed. Eng. 53 (10), 1943–1953.
Goldshtein, E., et al., 2011. Automatic detection of obstructive sleep apnea using speech signals. IEEE Trans. Biomed. Eng. 58 (5), 1373–1382.
Guariglia, E., 2017. Spectral analysis of the weierstrass-mandelbrot function. In: Proc. International Multidisciplinary Conference on Computer and Energy Science (SpliTech), Split, pp. 1–6.
Holmberg, E., Hillman, R., Perkell, J., 1988. Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. J. Acoust. Soc. Am. 84, 511–529.
Joseph, S.M., et al., 2009. Acute decompensated heart failure: contemporary medical management. Tex. Heart Inst. J. 36 (6), 510–520.
Keerthana, Y.M., Reddy, M.K., Rao, K.S., 2019. CWT-based approach for epoch extraction from telephone quality speech. IEEE Signal Process. Lett. 26 (8), 1107–1111.
Keras (online). Available at: https://github.com/fchollet/keras.
Kiran Reddy, M., Alku, P., Sreenivasa Rao, K., 2020. Detection of specific language impairment in children using glottal source features. IEEE Access 8, 15273–15279.
König, A., et al., 2015. Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease. Alzheimer's Dementia 1 (1), 112–124.
Mallat, S.G., 1989. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. 11 (7), 674–693.
Maor, E., et al., 2016. The sound of atherosclerosis: voice signal characteristics are independently associated with coronary artery disease. Circulation 134.Suppl_1, pp. A15840-A15840
Mirsamadi, S., et al., 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2227–2231.
Murton, O.M., et al., 2017. Acoustic speech analysis of patients with decompensated heart failure: a pilot study. J. Acoust. Soc. Am. 142 (4), EL401–EL407.
Orozco-Arroyave, J.R., et al., 2015. Characterization methods for the detection of multiple voice disorders: neurological, functional, and laryngeal diseases. IEEE J. Biomed. Health Inform. 19 (6), 1820–1828.
Pedregosa, F., et al., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.
Ponikowski, P., et al., 2016. ESC guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur. Heart J. 37 (27).2129–200
Rector, T.S., et al., 2006. Relationships between clinical assessments and patients' perceptions of the effects of heart failure on their quality of life. J. Card. Fail. 12 (2), 87–92.
Saenz-Lechon, N., et al., 2006. Methodological issues in the development of automatic systems for voice pathology detection. Biomed. Signal Process. Control 1 (2), 120–128.
Sandri, M., Zuccolotto, P., 2008. A bias correction algorithm for the gini variable importance measure in classification trees. J. Comput. Graph. Stat. 17 (3), 611–628.
Shen, P., et al., 2017. Conditional generative adversarial nets classifier for spoken language identification. In: Proc. INTERSPEECH, Stockholm, Sweden, pp. 2814–2818.
Zambroski, C.H., et al., 2005. Impact of symptom prevalence and symptom burden on quality of life in patients with heart failure. Eur. J. Cardiovasc. Nurs. 4 (3), 198–206.
Zheng, X., Tang, Y.Y., Zhou, J., 2019. A framework of adaptive multiscale wavelet decomposition for signals on undirected graphs. IEEE Trans. Signal Process. 67 (7), 1696–1711.
Ziaeian, B., Fonarow, G.C., 2016. Epidemiology and aetiology of heart failure. Nat. Rev. Cardiol. 13 (6), 368–378.