VocDoc, what happened to my voice? Towards automatically capturing vocal fatigue in the wild

Florian B. Pokorny, Julian Linke, Nico Seddiki, Simon Lohrmann, Claus Gerstenberger, Katja Haspl, Marlies Feiner, Florian Eyben, Martin Hagmüller, Barbara Schuppler, Gernot Kubin, Markus Gugatschka

Angaben zur Veröffentlichung / Publication details:

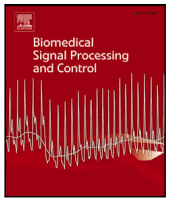
Pokorny, Florian B., Julian Linke, Nico Seddiki, Simon Lohrmann, Claus Gerstenberger, Katja Haspl, Marlies Feiner, et al. 2023. "VocDoc, what happened to my voice? Towards automatically capturing vocal fatigue in the wild." *Biomedical Signal Processing and Control* 88 (B): 105595. https://doi.org/10.1016/j.bspc.2023.105595.

Biomedical Signal Processing and Control 88 (2024) 105595

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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



VocDoc, what happened to my voice? Towards automatically capturing vocal fatigue in the wild

Florian B. Pokorny [a,b,c,1,*], Julian Linke [d,1], Nico Seddiki [d], Simon Lohrmann [d], Claus Gerstenberger [a], Katja Haspl [a], Marlies Feiner [a], Florian Eyben [e], Martin Hagmüller [d], Barbara Schuppler [d], Gernot Kubin [d], Markus Gugatschka [a]

[a] Division of Phoniatrics, Medical University of Graz, Austria
[b] EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[c] Center for Interdisciplinary Health Research, University of Augsburg, Germany
[d] Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
[e] audEERING GmbH, Gilching, Germany

ARTICLE INFO

Keywords:
Vocal fatigue
Voice features
Voice assessment
Speech-language pathology
Machine learning
Mobile application
Digital health

ABSTRACT

Objective: Voice problems that arise during everyday vocal use can hardly be captured by standard outpatient voice assessments. In preparation for a digital health application to automatically assess longitudinal voice data 'in the wild' – the *VocDoc*, the aim of this paper was to study vocal fatigue from the speaker's perspective, the healthcare professional's perspective, and the 'machine's' perspective.
Methods: We collected data of four voice healthy speakers completing a 90-min reading task. Every 10 min the speakers were asked about subjective voice characteristics. Then, we elaborated on the task of elapsed speaking time recognition: We carried out listening experiments with speech and language therapists and employed random forests on the basis of extracted acoustic features. We validated our models speaker-dependently and speaker-independently and analysed underlying feature importances. For an additional, clinical application-oriented scenario, we extended our dataset for lecture recordings of another two speakers.
Results: Self- and expert-assessments were not consistent. With mean F1 scores up to 0.78, automatic elapsed speaking time recognition worked reliably in the speaker-dependent scenario only. A small set of acoustic features – other than features previously reported to reflect vocal fatigue – was found to universally describe long-term variations of the voice.
Conclusion: Vocal fatigue seems to have individual effects across different speakers. Machine learning has the potential to automatically detect and characterise vocal changes over time.
Significance: Our study provides technical underpinnings for a future mobile solution to objectively capture pathological long-term voice variations in everyday life settings and make them clinically accessible.

1. Introduction

Speech is one of the most important means of human communication. A healthy voice as the underlying 'tool' for speech production, therefore, represents a fundamental requirement for social life as well as for a wide range of professions including, among others, teacher, lecturer, actor, singer, politician, reporter, call centre agent, guide, and priest. Voice problems, such as chronic hoarseness, may result from excessive voice use. They can strongly influence quality of life and job function [1], and even lead to social isolation and occupational disability [2,3]. An early detection of voice-related symptoms and a targeted therapy are thus essential.

1.1. State of the art

A commonly used voice evaluation protocol was released by the European Laryngological Society (ELS). It defines the minimum requirements for an objective functional voice assessment needed to diagnose common voice disorders and to follow up related therapy outcomes in an outpatient setting [4–6]. To this end, the protocol specifies norm categories and norm values for five non-redundant dimensions of assessment [4,5]: perception, video stroboscopy, aerodynamics, acoustics, and self-evaluation. Obviously, the protocol is very elaborate and can

* Corresponding author at: Division of Phoniatrics, Medical University of Graz, Austria.
E-mail address: florian.pokorny@medunigraz.at (F.B. Pokorny).
[1] Contributed equally.

https://doi.org/10.1016/j.bspc.2023.105595
Received 20 July 2023; Received in revised form 4 October 2023; Accepted 14 October 2023
Available online 25 October 2023

be performed in a clinical setting only, e. g., by a speech and language therapist (SLT). Another inherent limitation of an outpatient voice assessment is that it only provides a snapshot of the patient's voice, which is usually in a relaxed state. Therefore, a vocal stress test was established by Schneider-Stickler and Bigenzahn [7], but it is rarely used. The test is computer-aided and evaluates vocal endurance and resilience. The patient is asked to phonate over a period of approximately 20 min at a certain minimum sound level. A microphone is used to record the phonation sequence for the analysis of acoustic basic features, such as the fundamental frequency (f_o), jitter, and shimmer. An optimal vocal endurance is given if the patient completes the test without audible changes in voice quality, without changes in acoustic periodicity features including sound level-related f_o dynamics, and without laryngeal morphological changes as compared to the beginning of the test. An initial increase of the f_o until the patient reaches the required minimum sound level is regarded as physiological. However, a further increase of the f_o throughout the test while phonating at a constant sound level is an indicator of vocal compensation efforts due to vocal fatigue [7].

Even tough a vocal stress test simulates to some extent the long-term use of the voice and, thus, represents a useful 'add-on tool' to the ELS basic voice evaluation protocol, it does not necessarily reflect the natural use of the voice in everyday life settings. Therefore, subjective vocal discomfort and perceived effects on voice function that emerge after a certain longer period of voice use – reported by patients, who need to talk for many hours a day for occupational reasons – can usually not be clinically captured. This hampers optimal treatment.

Catalysed by recent technological advancements – especially in sensor and communication technology, as well as in machine learning – a paradigm shift towards personalised and connected medicine is currently taking place [8]. Vital parameters, as well as behaviour-related data, such as heart rate, respiration rate, blood pressure, blood oxygen saturation, body temperature, sleep phases, activity pattern, walked steps, and GPS information, captured over the day (and night) by means of wearable devices are playing an increasingly important role in medical research in context of both physical and mental health [9–11]. Consequently, high expectations are currently associated with remote assessment approaches in addition to clinical on-site assessments to improve future diagnostic procedures, intervention paradigms, as well as disease prevention strategies.

In line with this trend, we suggest a smartphone-based solution to objectively capture long-term variations of the voice in everyday life. Previous studies mainly reported changes in a few single voice features only, such as a rise in f_o, sound pressure level, shimmer, or noise-to-harmonics ratio, coming along with prolonged vocal loading [12–18].

1.2. Contribution

In this work, we aimed to study vocal fatigue in voice healthy individuals from different, mutually related perspectives, namely

(i) the speaker's perspective (self-assessment),
(ii) the perspective of healthcare professionals (expert assessment), and
(iii) the audio signal processing and machine learning perspective (machine assessment).

For (ii) and (iii), we elaborated on the task of elapsed speaking time recognition, similarly as done by Bayerl et al. [19]. On the basis of an extended set of acoustic voice features, we further aimed to find out whether there is a subset of features well suited for capturing vocal fatigue across different speakers and, thus, for a population-based vocal fatigue classification approach. Thereby, this study shall provide an empirical foundation for the realisation of a mobile application to capture clinically relevant long-term voice variations 'in the wild'. A prototype of a long-term voice recording and feature extraction app – the *VocDoc* – has already been developed in collaboration with audEERING GmbH, Germany. In a next step, methods implemented in the prototype shall be optimised on the basis of knowledge gained in this study and tested in a clinical real-world scenario.

Table 1
Dataset overview in terms of gender (f = female; m = male), age, number of (#) recordings, and recording duration per speaker. (1)/(2) = with regard to first/second lecture recording.

Subset	Speaker ID	Gender	Age [years]	# Recordings	Duration [min]
pilotSet	101M	m	30	1	90
pilotSet	102M	m	31	1	90
pilotSet	103M	m	30	1	90
pilotSet	104M	m	50	1	90
extension	111M	m	49	2	(1) 91; (2) 95
extension	114F	f	(1) 28; (2) 29	2	(1) 99; (2) 91

2. Materials and methods

Data collection and experimentation were carried out at the Graz University of Technology and the Medical University of Graz, Austria.

2.1. Data collection

Experiments in this study were based on the *VocDoc-pilotSet*, which consists of speech data from a simulated lecture setting. Later on, we combined the *VocDoc-pilotSet* with a set of real-world lecture recordings, referred to as *VocDoc-pilotSet extension*.

2.1.1. VocDoc-pilotSet

The *VocDoc-pilotSet* comprises recordings from four male speakers (authors JL, NS, SL, and CG) staged in a lecture hall at the Graz University of Technology, to simulate presentation talks (see Table 1). The speakers' first language was German. None of them was a trained lecturer. Voice health of the speakers was verified in advance by means of endoscopy at the Division of Phoniatrics, Medical University of Graz. The speech material was recorded with a Tascam DRX-05 stereo field recorder, where each speaker was standing in front of the microphone at a distance of approximately 1 m. All speakers read out loud the same text from the German scientific book "Sprachverarbeitung – Grundlagen und Methoden der Sprachsynthese und Spracherkennung" (Engl.: "Speech processing – Fundamentals and methods of speech synthesis and speech recognition") by Pfister and Kaufmann [20]. Speakers were ending their readings at different stages of the book, since we restricted each talk to exactly 90 min. This led to 6 h of speech data in total.

2.1.2. VocDoc-pilotSet extension

The *VocDoc-pilotSet extension* consists of 4 recordings of real lectures of approximately 90 min provided by the internal video portal of Graz University of Technology.[2] Two of the lectures were held by the same male lecturer, the other two by the same female lecturer (see Table 1). Both lecturers were native German speakers and the lectures were also held in German. Informed consent was obtained from the lecturers for the analysis of their voices in the framework of this study.

2.2. Self-assessment

Immediately before we started to record each session of the *VocDoc-pilotSet*, we asked the speakers, if they had already stressed their voice at that day (**Q1**). In order to find out if, how, and at which point in time the subjective perception of speaking had changed, we then interrupted the reading every 10 min and performed a question & answer session of approximately 2 min. The questionnaire involved the following questions/tasks:

[2] https://tube.tugraz.at/.

Q2: *How did the voice feel in the last 10 min?*
Q3: *How does your voice feel right now?*
Q4: *How are you feeling right now in general?*
Q5: *Do you feel you have spoken fast in the last 10 min?*
Q6: *Mark the location(s) where you feel vocal strain. (Given an illustration of a cut of the vocal tract.)*

Q2, **Q3**, and **Q4** included 3 answer options, respectively; **Q5** included one answer option relating to the speech rate. In case of **Q2** and **Q3**, the 3 answer options referred to stress, fatigue and roughness. In case of **Q4**, the 3 answer options referred to excitement, alertness and concentration. For each category, the speakers could (verbally) choose an integer value from 0 to 10, with 0 standing for "not applicable at all" and 10 standing for "entirely applicable". Finally, for the task **Q6**, participants could (verbally) mark potentially stressed regions based on a presented illustration of a midsagittal cut of the vocal tract, such as nasal cavity, alveolar ridge, tongue, palate, velum, pharynx, or larynx.

For the present study, we focused on the answers to **Q3** and **Q4** only, since – in our preliminary approach – we preferred an exact time resolution (answers referring to the current status and not to a certain, not specified point in time or period within a 10-min block). Moreover, the answers to **Q6** were saved for specific future work on vocal fatigue in relation to voice anatomy/physiology.

2.3. Expert assessment

To consider the perspective of clinicians, who work with patients with voice problems on a daily basis, we asked two SLTs (authors KH and MF) from the Department of Otorhinolaryngology, University Hospital Graz, Austria, to perceptually evaluate the recordings of the four speakers from the *VocDoc-pilotSet*. Specifically, we wanted to find out, if these experts are able to recognise elapsed speaking time when quantised into three periods. For this purpose, we extracted exactly one consecutive minute of speech from (i — start) the first 10 min, (ii — mid) the interval from 40 to 50 min, and (iii — end) the last 10 min (i. e., the interval from 80 to 90 min) of each speaker's recording. As all speakers read aloud the same book starting at chapter 1 and stopping at different places due to individual reading rates over 90 min, we decided to select continuous one-minute start, mid, and end segments for the expert assessment, that do not overlap in terms of content. Moreover, we only selected segments, which do not contain any read out chapter, figure, or table numbers, as they would have been cues for reading progress. The SLTs were asked to independently listen to the three one-minute audio segments of each speaker in a quiet room using studio headphones. They were allowed to listen to each speaker's clips as often as desired and they could jump between clips of the same speaker as well as jump forth and back within a single clip. In doing so, the experts had the task to put the clips of each speaker in correct order, i. e., start – mid – end. In addition, they were asked to note down which speech or voice attributes led them to their decisions.

2.4. Machine assessment

Alongside the human rater-based methods, i. e., the self-assessment and the expert assessment, we carried out machine learning experiments. In doing so, we extracted acoustic features from the recordings (see Section 2.4.1), investigated tasks of elapsed speaking time classification on a block-wise basis (see Section 2.4.2), and, thereby, derived acoustic features that best characterise potential effects of vocal fatigue across different speakers (see Section 2.4.3).

2.4.1. Acoustic feature extraction

Given the self-assessment design of interrupting the speakers of the *VocDoc-pilotSet* on a 10-min basis for a short questioning, we cut each recording into 9 10-min blocks that exclusively contain reading (and not the questioning). Equally, we also split each recording of the *VocDoc-pilotSet extension* into 9 10-min blocks and discarded audio exceeding the 90 min mark (see Table 1). Subsequently, we applied the widely-used open-source toolkit openSMILE [21,22] (version 3.0.1) to all 10-min blocks to (i) automatically segment the material into utterances by means of the included pre-trained long short-term memory recurrent neural network (LSTM-RNN) based voice activity detector [23], and to (ii) subsequently extract acoustic features from each utterance. openSMILE is implemented in C++ and uses a ring-buffer memory and a modular architecture, which allows for arbitrary combinations of audio signal processing and calculation steps set in a single configuration file. Automatic segmentation led to 100–170 utterances of 1–7 s per 10-min block. For the feature extraction step, we used the most current standardised openSMILE feature set, i. e., the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [24]. The eGeMAPS represents the openSMILE set of choice for baseline evaluations in computational voice/speech analysis tasks – including clinically-related tasks – when aiming for compactness and feature interpretability at reasonable machine performance. The set comprises 88 acoustic features, most of which are statistical functionals (such as arithmetic mean or coefficient of variation) applied to smoothed trajectories of frequency-, energy/amplitude-, and spectrum-related low-level descriptors (such as f_0, loudness, or the Hammarberg index).

For subsequent classification experiments, the original continuous audio recordings were, thus, transferred into 88-dimensional numeric feature vectors acoustically representing single utterances assigned to specific 10-min blocks. For non-commercial research purposes, all feature files are available upon request from the corresponding author.

2.4.2. Classification

In order to investigate if and after which period of vocal strain the voice changes on the basis of 10-min time blocks, we examined the problem of automatic elapsed speaking time recognition translated into a series of 8 binary classification tasks, where each model was trained to differentiate between utterances from the first 10-min block (class **Ref**) and utterances from 1 of the 8 successively adjacent 10-min blocks (set of classes $\{\mathbf{S}_1, \ldots, \mathbf{S}_8\}$), respectively (see Fig. 1). For reasons of better explainability in terms of relevance of underlying features (see Section 2.4.3) as compared to deep neural networks, we trained each classification model with a random forest (RF) by utilising the scikit learn toolkit (version 1.0.1) [25] developed for Python (3.9.7). The number of decision trees per RF model was set to 100 and the quality of a split was measured with the Gini criterion. Further, we chose a maximum depth of 20 nodes and a minimum split of 10 samples. Experiments were carried out on a Windows 10 Pro PC with an Intel Core i7-7600 2.80 GHz (2 Cores) CPU.

We conditioned our experiments on three training and validation scenarios (see also Table 2):

- Speaker-dependent (SD) classification: eight RF models utilising the *VocDoc-pilotSet*; training and validation in separate for all four speakers.
- Speaker-independent (SI) classification: eight RF models again utilising the *VocDoc-pilotSet*, but now with leave-one-speaker-out cross-validation; training on three speakers and testing on the left-out speaker per validation run with each of the four speakers used for testing exactly one time.
- Speaker-independent personalised (SIP) classification: eight RF models utilising the *VocDoc-pilotSet* plus the *VocDoc-pilotSet extension*; training and validation in separate for the two speakers (one female and one male) from the *VocDoc-pilotSet extension* with training on all four speakers from the *VocDoc-pilotSet* plus the first

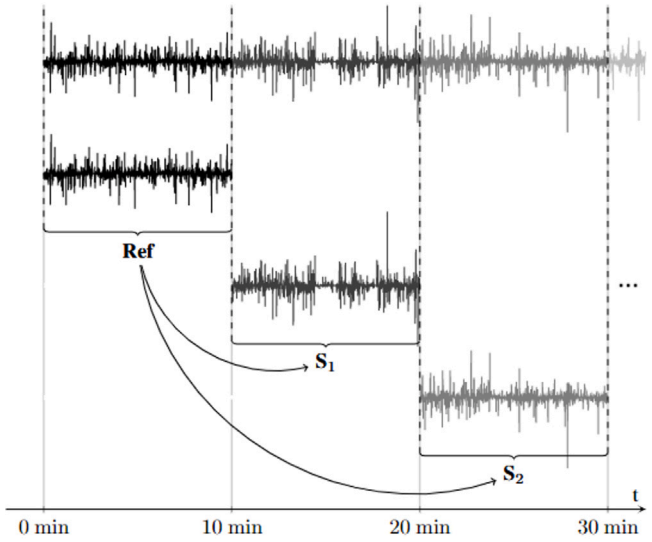


Fig. 1. Classification scheme based on 10-min blocks. A series of binary classification models were trained to distinguish between the reference block, i. e., the first 10-min block (class **Ref**), and successively adjacent 10-min blocks (set of classes $\{\mathbf{S}_1, \mathbf{S}_2, \ldots\}$).

Table 2
Training and test sets per scenario and validation run with respect to speaker IDs. The *VocDoc-pilotSet* comprises the speakers 101M–104M; the *VocDoc-pilotSet extension* comprises the speakers 111M and 114F. SD = speaker-dependent, SI = speaker-independent, SIP = speaker-independent personalised; M and F in speaker IDs encode speaker gender with M = male and F = female.

Scenario	Training	Test
SD	101M	101M
SD	102M	102M
SD	103M	103M
SD	104M	104M
SI	{102M, 103M, 104M}	101M
SI	{101M, 103M, 104M}	102M
SI	{101M, 102M, 104M}	103M
SI	{101M, 102M, 103M}	104M
SIP	{101M, 102M, 103M, 104M}+**111M**	**111M**
SIP	{101M, 102M, 103M, 104M}+**114F**	**114F**

recording of the selected speaker from the *VocDoc-pilotSet extension* and testing on the second recording of the (same) selected speaker from the *VocDoc-pilotSet extension*.

The SIP scenario was chosen to examine the potential future application of equipping an individual (patient) with a smartphone app with a pre-trained population-based vocal fatigue recognition model that gradually adapts to the user by retraining based on data collected from the individual over a certain initial period.

Performance evaluation. In order to evaluate the performance of our classification experiments we introduce a broad comparison of resulting harmonic means, i. e., F1 scores, which were calculated from both, the precisions and recalls of a given test result:

$$\text{F1 score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{1}$$

Since our experiments were conditioned on three scenarios with test speakers (i) present in the training data (SD/SIP experiments) or (ii) not present in the training data (SI experiments), our evaluations of the SD/SIP experiments were slightly different from the evaluations of the SI experiments. In case of the SD/SIP experiments, we compared speaker-dependent F1 scores by generating only one random training/test-split with a ratio of 75%/25% of time for the reference block and several training/test-splits for each shifted block. This was done by performing a minute-wise 10-fold cross-validation with respect to the shifted block only. Thus, for each comparison between the fixed reference block and a shifted block, we tested 10 different 1-minute-intervals of the shifted block by including remaining data in the training splits, respectively. In case of the SI experiments we again calculated F1 scores, but this time without performing within-block-cross-validation, leading to 20 min of available test data for each binary classification.

For each scenario, we analysed our initially generated models for the respective ten most relevant acoustic features (see Section 2.4.3). As we found that classification performance slightly increased when retraining the models just based on the identified top ten features, we present results for those retrained models only.

2.4.3. Acoustic feature analysis

Our choice of RFs for classification allowed us to examine feature importances for each 10-min block. These importances reveal the worth of the different features with respect to capturing elapsed speaking time and, thus, implicitly potential effects of vocal fatigue. We exploited the RF feature importances resulting from the SD, SI, and SIP experiments in order to select the respective best global features. Identifying the best global features was done as follows: First, for each experiment we extracted the feature importances for every 10-min block. Second, all features were summed across all time blocks for one speaker (SD/SIP experiments) and over all speakers (SI experiments).

In general, our method resulted in $N_F \times N_T \times N_S$ feature importances, where N_F is the total number of features, N_T is the respective number of RF models, and N_S is the respective number of speakers. For each validation scenario, we summarised the features in a feature importance matrix, in which the number of rows corresponded to the number of features N_F and the number of columns corresponded to the number of 10-min blocks (= number of RF models N_T). Thus, the resulting matrix for one speaker had a dimension of $N_F \times N_T$. We defined the matrix as $\mathbf{I}_s = [\mathbf{i}_{s,1}, \mathbf{i}_{s,2}, \mathbf{i}_{s,3}, \ldots, \mathbf{i}_{s,N_T}]^T$ with $\mathbf{i}_{s,1\ldots N_T} \in \mathrm{R}^{N_F \times 1}$. In order to summarise feature importances for a specific speaker with respect to all time blocks, we summed over the columns of $\mathbf{I}_s$ and normalised by the number of models N_T. Finally, we calculated normalised global feature importances by summing over resulting summarised importances of all speakers and normalised with respect to N_S. This can be described with the following formulae:

$$\hat{\mathbf{i}} = \frac{1}{N_s} \sum_{s=1}^{N_s} \frac{1}{N_T} \sum_{t=1}^{N_T} \mathbf{i}_{s,t} = \frac{1}{N_s \cdot N_T} \sum_{s=1}^{N_s} \sum_{t=1}^{N_T} \mathbf{i}_{s,t}, \tag{2}$$

where $\hat{\mathbf{i}} \in \mathrm{R}^{N_F \times 1}$ contains the averaged feature importances for each validation scenario. Assuming that the vector components are already sorted in descending order, we defined a set where the elements were the ordered vector components of

$$\mathrm{I} = \{i_1, \ldots, i_{N_F}\}. \tag{3}$$

Defining M as the number of highest components and defining

$$\mathrm{I}_{(-M)} = \max(m : \#\{i \in \mathrm{I} : i \geq m\} = M) \tag{4}$$

gave us

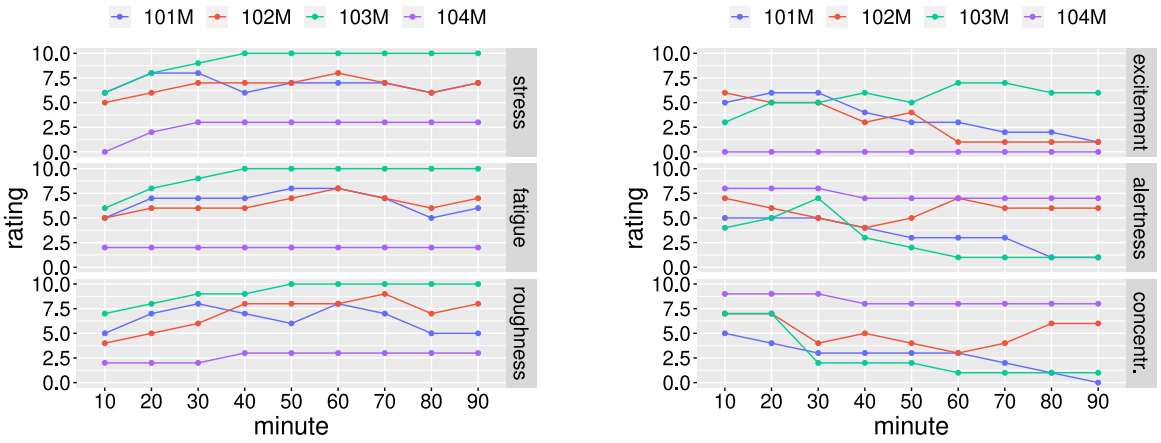
$$\mathrm{I}^{(M)} = \{i \in \mathrm{I} : i \leq \mathrm{I}_{(-M)}\}. \tag{5}$$

For each validation scenario (SD, SI and SIP), we report the respective combination of $M = 10$ best features.

3. Results

3.1. Self-assessment

Speaker self-assessment results for **Q3** and **Q4** of the questionnaire (see Section 2.2) are summarised in Figs. 2(a) and 2(b).



(a) How does your voice feel right now?

(b) How are you feeling right now in general?

Fig. 2. Speaker-wise (101M–104M) questionnaire ratings (from 0 = not applicable at all, to 10 = entirely applicable) for questions (a) **Q3** and (b) **Q4**. Concentr. = concentration.

With regard to **Q3**, which relates to stress, fatigue and roughness, we observed generally higher ratings in speakers 101M, 102M and 103M as compared to speaker 104M. Ratings of speakers 101M and 102M showed a similar trajectory with values ranging between 4 and 9. Ratings of speaker 103M increased right from the start reaching the maximum scale value (10) after 40 min for stress and fatigue, and after 50 min for roughness. Those ratings did not change any more till the end of the recording.

With regard to **Q4**, which relates to excitement, alertness and concentration, we observed a contrary picture. The excitement ratings of speaker 101M started at 5, slightly increased to 6 for minutes 20 and 30, and then decreased continuously over time till a rating of 1 in minute 90. Speaker 102M had similar excitement ratings starting at 6 but decreasing to 1 already after 60 min for the rest of the recording. In speaker 103M, the excitement rating increased a bit over time starting at 3 and ending at 6 with minor fluctuations in between. There was no change in the excitement rating of speaker 104M, who gave a rating of 0 at each questioning. With respect to the alertness ratings, speakers 101M and 103M were starting at 5 and 4, respectively, and decreasing to 1 over time with one distinct peak of 7 in speaker 103M at minute 30. In speaker 102M, we observed an opposite behaviour since ratings started at a score of 7, then decreased to 4 at minute 40, increased again to 7 at minute 60, and finally ended in a score of 6 from minute 70 onward until the end. Speaker 104M again gave almost constant ratings starting at a score of 8 and ending at a score of 7. The concentration ratings of speakers 101M (starting at a rating of 5) and 103M (starting at a rating of 7) decreased over time and ended at a score of 1 in case of speaker 103M (minutes 60–90) and a score of 0 in case of speaker 101M. Corresponding ratings of speaker 102M started at 7, declined to 3–5 for minutes 30–70 and, finally, rose again to a score of 6 at the last two questionings. Speaker 104M rated concentration similar to alertness, but this time starting at a score of 9 and ending at a score of 8.

Globally, there was a tendency of increasing vocal stress, fatigue, and roughness and decreasing general excitement, alertness and concentration with increasing speaking time.

3.2. Expert assessment

Changing over from the speaker's subjective perspective to the objective perspective of healthcare professionals experienced in perceptual voice assessment, the performance of the two SLTs in recognising whether a speaker (i) had just started to speak, (ii) already had spoken for more than 40 min, or (iii) already had spoken for more than 80 min, is presented in Table 3. The first expert managed to correctly identify

Table 3
Confusion matrix showing the speaker-wise segment order ratings by expert 1 (upper left) and expert 2 (lower right) in each cell. Green = correct segment order, red = incorrect segment order.

Speaker	Segment (S)tart	Segment (M)id	Segment (E)nd
101M	M / S	S / E	E / M
102M	S / S	M / E	E / M
103M	S / S	M / M	E / E
104M	S / E	M / S	E / M

the order of segments in three of the four speakers (102M–104M) as well as the end segment of the remaining speaker (101M) with the start and the mid segment of this speaker having been mixed up (see upper left halves of cells in Table 3). Reported attributes that were related to the first expert's correct decisions were an increase in the f_o (102M), reading rate (102M), occurrence of swallowing and harrumphing events (102M), and vocal fry (103M), a decrease in the fraction of chest voice (104M), clarity (102M, 103M), loudness (103M), and fluency (104M) of speaking, as well as a change in intonation at the end of sentences from more to less melodic (102M), over time. In contrast, the second expert correctly identified the order of segments in one of the four speakers (103M), as well as the respective start segments of two speakers (101M and 102M) with the mid and the end segments of these speakers having been mixed up (see lower right halves of cells in Table 3). For the correctly rated start segments of speakers 101M and 102M, the second expert reported a slower/more comfortable reading rate as compared to the other segments. Identified attributes of increasing speaking time in speaker 103M were an increase in the f_o, reading errors, voice creakiness, mouth dryness, phonation pressure, and pause duration. Not a single segment of the remaining speaker (104M) was correctly identified by the second expert, who described this speaker's voice and reading performance as rather constant over time with just minor changes in phonation pressure and reading rate. When converting the original expert assessment paradigm into a binary recognition task by discarding the mid segment, results show that it happened just one time that the order start–end was not correctly identified, namely by the second expert in speaker 104M.

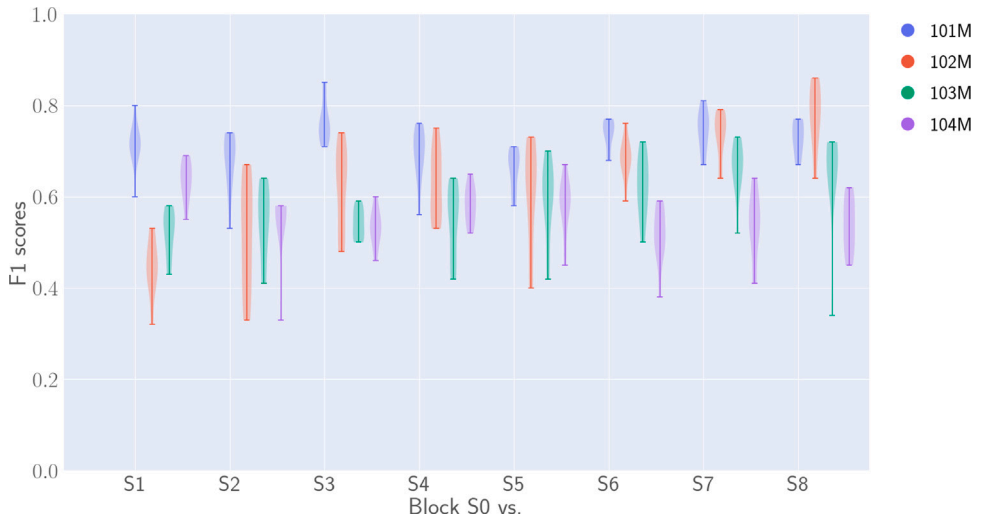


Fig. 3. Speaker-wise (101M–104M) distribution of F1 scores for each block derived from the speaker-dependent experiments.

3.3. Machine assessment

Finally, we present our machine assessment results. Fig. 3 summarises F1 score distributions derived from the SD experiments.

Across the first classification models, which distinguish between **Ref** and $\mathbf{S}_1$, we achieved the best results for speakers 101M (mean F1 score: 0.71) and 104M (mean F1 score: 0.64). In general, mean F1 scores of speaker 101M were between 0.66 and 0.76 and did not change much over time. In contrast, speakers 102M, 103M and 104M had more varying F1 scores. In case of speaker 102M, mean F1 scores improved from 0.45 (block $\mathbf{S}_1$) to 0.78 (last block). Mean F1 scores of speaker 103M slightly improved from 0.52 (block $\mathbf{S}_1$) to 0.65 (block $\mathbf{S}_8$). In speaker 104M, we observed a different behaviour. We found a decrease of mean F1 scores at the beginning from 0.64 (block $\mathbf{S}_1$) to 0.55 (block $\mathbf{S}_2$). Then, for all subsequent blocks, mean F1 scores did not vary a lot over time fluctuating between 0.52 and 0.58.

Regarding feature importances, we observed that for each individual speaker there were specific features among the global top ten features which were more relevant than others. Respective distributions are shown in Fig. 4. In addition, we present correlation plots of each speaker's respective most relevant feature across time blocks in Fig. 5 and reveal Spearman's correlation coefficients (ρ). In speaker 101M, the mean energy ratio between the first and the second f_0 harmonic was the most important feature with a proportion $\geq$0.3 in 6 blocks, and the coefficient of variation (standard deviation normalised by the arithmetic mean) of the same energy ratio was the second most important feature with proportions $\geq$0.17 again in 6 blocks. Additionally, the mean spectral flux in voiced regions turned out important in blocks $\mathbf{S}_5, \mathbf{S}_6, \mathbf{S}_7$, and $\mathbf{S}_8$. In speaker 102M, the mean second Mel-frequency cepstral coefficient ($MFCC_2$) in voiced regions and the mean $MFCC_2$ in the entire segment were most important with respective proportions $\geq$0.14 and $\geq$0.11 across all blocks. A weak correlation ($\rho = 0.35$) between the mean $MFCC_2$ in voiced regions and the block index was found. Blocks $\mathbf{S}_4$ and $\mathbf{S}_5$ indicated a higher importance of the mean Hammarberg index in voiced regions (proportion $\geq$0.17). For the same speaker, the mean spectral flux in voiced regions resulted in the highest proportion of 0.24 in the last block. Feature importances of speaker 103M gave highest proportions $\geq$0.16 for the mean first formant frequency at a weak negative correlation with the block index of $\rho = -0.20$. In speaker 104M, proportions of feature importances indicated that the mean spectral slope from 0–500 Hz in voiced regions, the mean spectral slope from 0–500 Hz in unvoiced regions, as well as the mean spectral flux in voiced regions had the highest impact on classification results. Additionally, the mean third formant bandwidth was the most important feature for block $\mathbf{S}_5$.

In order to evaluate the generalisation capabilities towards new speakers, we present SI classification results. Again, we examined block-wise classification performances – here, for each speaker having been the test speaker exactly one time – and compared respective feature importances. For better (visual) comparability with the SD and SIP models, we split the test data from the respective shifted block again into ten parts and present F1 score distributions for the SI experiments in Fig. 6. Over all blocks and speakers, mean F1 scores only ranged between 0.22 and 0.36.

The importances of the underlying top ten features were more uniformly distributed than the top ten features of the SD scenario (see Fig. 7).

Finally, Fig. 8 reveals the classification performance under SIP validation, i. e., the scenario including the two additional speakers (111M and 114F) from the *VocDoc-pilotSet extension* for personalised training and testing.

Mean F1 scores and standard deviations of speakers 111M and 114F were widely homogeneous over time. Mean F1 scores ranged between 0.37 and 0.51 with standard deviations between 0.09 and 0.15 in the male speaker 111M, and between 0.45 and 0.54 with standard deviations between 0.06 and 0.13 in the female speaker 114F. In speaker 111M, mean F1 scores slightly increased subsequent to the first 4 blocks with a maximum of 0.42 for block $\mathbf{S}_6$. In speaker 114F, we observed a similar characteristic with the highest mean F1 score of 0.47 for block $\mathbf{S}_6$ as well.

Similar as for the SI scenario, we did not observe a single out of the scenario-specific top ten features that was clearly more important than the others (see Fig. 9).

4. Discussion

The development of a mobile solution for the automatic detection and characterisation of long-term variations of the voice in everyday life settings requires a profound understanding of the (patho-)physiology of vocal fatigue. In this study, we thus investigated vocal fatigue in healthy individuals completing a 90-min reading task. We provided results from a series of technical feasibility experiments on the automatic recognition of elapsed speaking time, complemented by insights into the speakers' subjective perception as well as SLTs' professional evaluation of vocal changes over time. In doing so, we tried to find out, (i) after which time and to which extent vocal fatigue affects speakers without any (known) a-priori voice problems, and (ii) whether vocal fatigue manifests itself in the same way with regard to acoustic features across different speakers.

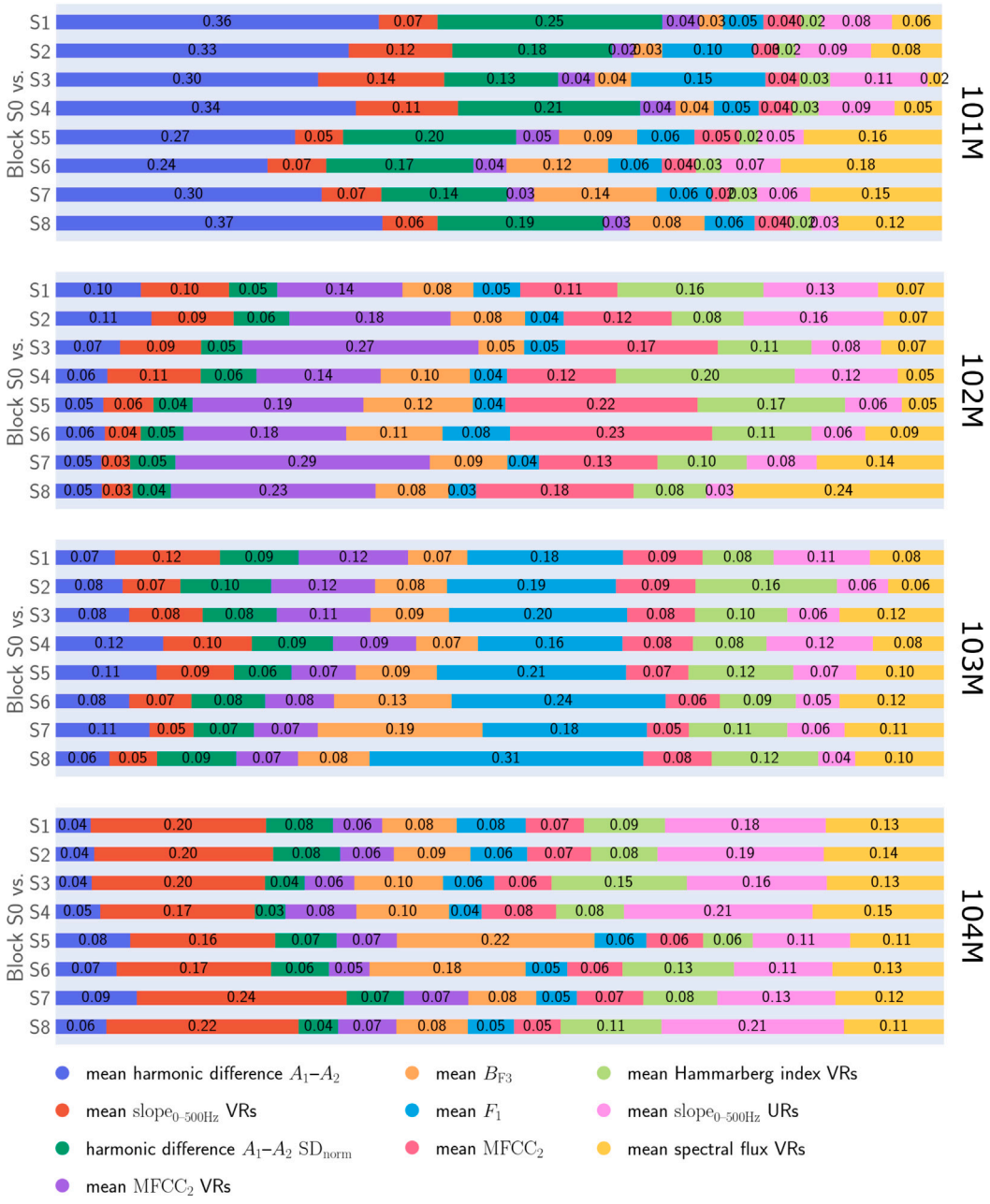


Fig. 4. Speaker-wise (101M–104M) random forest feature importances derived from the speaker-dependent experiments. $A_{1|2}$ = relative amplitude of first|second harmonic, B_{F3} = bandwidth of third vowel formant, F_1 = frequency of first vowel formant, MFCC_2 = second Mel-frequency cepstral coefficient, $\mathrm{SD}_{\mathrm{norm}}$ = standard deviation normalised by the arithmetic mean (coefficient of variation), URs = unvoiced regions, VRs = voiced regions.

We did not obtain a homogeneous picture: For speaker 102M, our SD experiments revealed that the more time elapsed, the better the automatic recognition of elapsed speaking time was. This might indicate that this speaker's voice continuously changed over time from the beginning onward with increasing extent. A similar tendency was found for speaker 103M as well. These findings are supported by the respective self-assessments as well as the ratings of the SLTs. Both experts correctly identified the order of segments in speaker 103M. In speaker 102M, the second expert just mixed up the mid and the end segment. Interestingly, in speaker 101M the best recognition performances were already achieved early on, which could mean that subsequent to initial voice changes compensation mechanisms became active. 101M was also the speaker, who caused the most incorrect ratings among both SLTs. However, this could also be an effect of rating chronology — both experts evaluated the recordings of the four speakers in order 101M, 102M, 103M, and 104M; thus, the experts might not have been entirely familiar with the task yet when evaluating the recordings of speaker 101M. In speaker 104M, the automatic elapsed speaking time recognition performance was widely constant and low over the entire 90-min period, which suggests that this speaker's voice did not change a lot over time. This is also in line with the self-ratings of this speaker, which mostly differed from the ratings of the other three speakers and were relatively constant over time. Moreover, speaker 104M was the only speaker, whose start, mid, and end segments were completely mixed-up by the second expert. At this point, it should be considered that speaker 104M has a significantly higher age (50 years) than the other speakers from the *VocDoc-pilotSet* ($\leq$ 31 years).

The poor results obtained in the SI experiments show that our models do not generalise well to other speakers. This might point out that vocal fatigue has individual effects across different speakers (and ages).

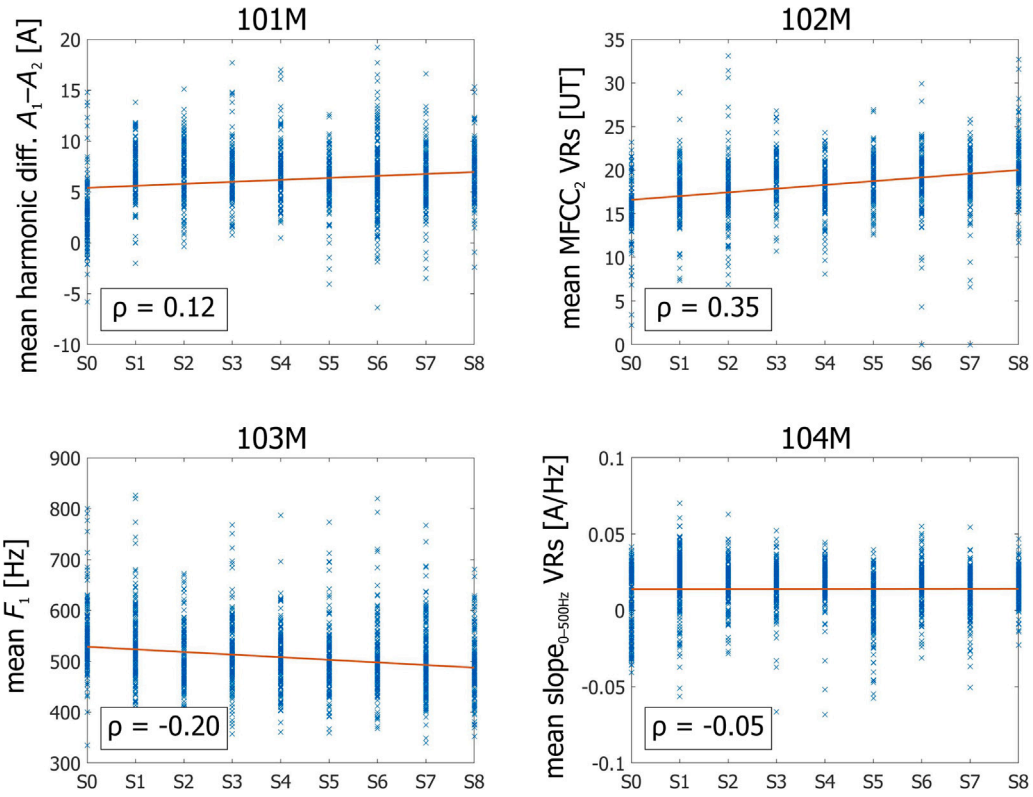


Fig. 5. Respective most relevant feature for each speaker (101M–104M) across time blocks (S0[(≙**Ref**)]–S8) and linear fitting line. Spearman's correlation coefficient (ρ) is given on the bottom left of each subfigure. A = amplitude, $A_{1|2}$ = relative amplitude of first|second harmonic, F_1 = frequency of first vowel formant, $MFCC_2$ = second Mel-frequency cepstral coefficient, UT = unit of time, VRs = voiced regions.

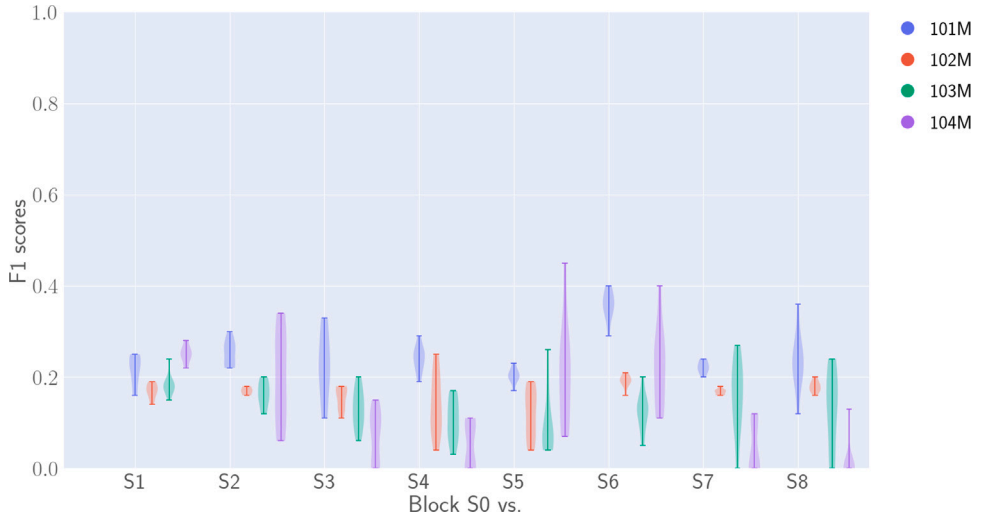


Fig. 6. Speaker-wise (101M–104M) distribution of F1 scores for each block derived from the speaker-independent experiments.

This finding is also supported by the results of our feature importance analyses. While the identified global top ten features of the SI scenario were rather uniformly distributed across all speakers and blocks at low elapsed speaking time recognition performance, there were speaker- and block-specific global top ten features of higher importance in combination with higher recognition performances in the SD scenario. In speaker 101M, the mean energy difference between the first and second harmonic – a measure reflecting glottal constriction [26] and, thus, associated with creaky and/or pressed phonation [26,27] – seems to bear relevant acoustic information related to vocal changes over speaking time. In speaker 102M, the most important features were the $MFCCs_2$ in voiced regions and in all regions, which can be regarded as weighted measures of a low to high frequency energy ratio, sensitive for increases in aspiration noise due to incomplete vocal fold closure [28]. MFCCs have generally shown great potential in a wide range of acoustic signal processing applications, such as in automatic speech recognition [29–32], speaker recognition [33,34], speech emotion recognition [35], speech synthesis [36], and speech coding [37]. Possible effects of vocal fatigue in speaker 103M were mainly reflected in the mean first formant frequency – a measure that is inversely related to vowel height. As the distribution of vowels is assumed to be constant over the entire read book (same text type, same language, same author), changes in the mean first formant frequency could indicate, that one and the same vowels were increasingly articulated with different qualities, thus, at

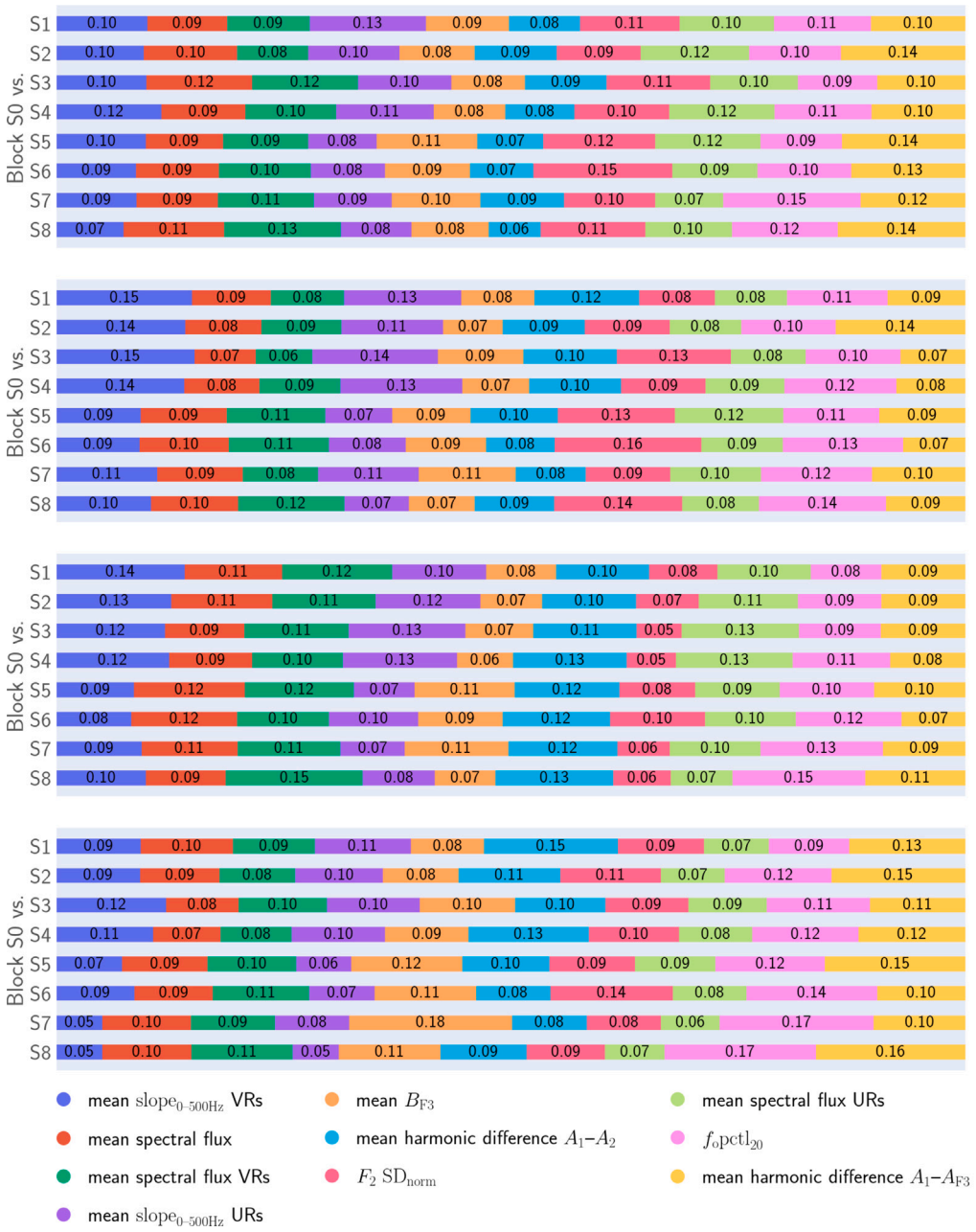


Fig. 7. Speaker-wise (101M–104M) random forest feature importances derived from the speaker-independent experiments. $A_{1|2}$ = relative amplitude of first|second harmonic, A_{F3} = amplitude of third vowel formant, B_{F3} = bandwidth of third vowel formant, f_o = fundamental frequency, F_2 = frequency of second vowel formant, pctl_{20} = 20th percentile, SD_{norm} = standard deviation normalised by the arithmetic mean (coefficient of variation), URs = unvoiced regions, VRs = voiced regions.

slightly shifted articulation points. In speaker 104M, the most relevant acoustic features were the mean spectral slopes from 0–500 Hz in voiced and unvoiced regions — voice quality measures that express the energy tilt in the audio spectrum towards higher frequencies within the specified frequency interval. Perceptually, these features are associated with the warmth of a male voice [27].

Most of the 88 eGeMAPS features turned out not to be that relevant for the acoustic characterisation of elapsed speaking time and, thereby, of potential effects of vocal fatigue, as they were not among the global top ten features of any validation scenario. In contrast, three features even appeared among the global top ten features across all scenarios. These were again the mean spectral slopes from 0–500 Hz in voiced and unvoiced regions, and the mean bandwidth of the third vowel formant, a measure that was also found to differ between healthy speakers and speakers with an acute COVID-19 infection when producing sustained back vowels /u:/ and /o:/ [38]. Previous studies reported on a rising f_o, sound pressure level, shimmer, and/or noise-to-harmonics ratio concomitant with a prolonged voice use [12–18]. In our experiments, not a single energy/amplitude-related feature, i. e., shimmer, harmonics-to-noise ratio, and loudness, made it into the global top ten features of any validation scenario. However, the mean f_o was among the most important features for the SIP scenario, the 20th percentile of the f_o among the most important features for both the SI and the SIP scenario.

The performances of our SIP models were in between the performances of the SD and the SI models, which is not surprising as we here applied a sort of speaker adaptation by combining population-based/speaker-independent and speaker-dependent training and validation [39]. Interestingly, there were no clear performance differences between the models for the male and the female speaker, even though all speaker-independent training data stemmed from male speakers.

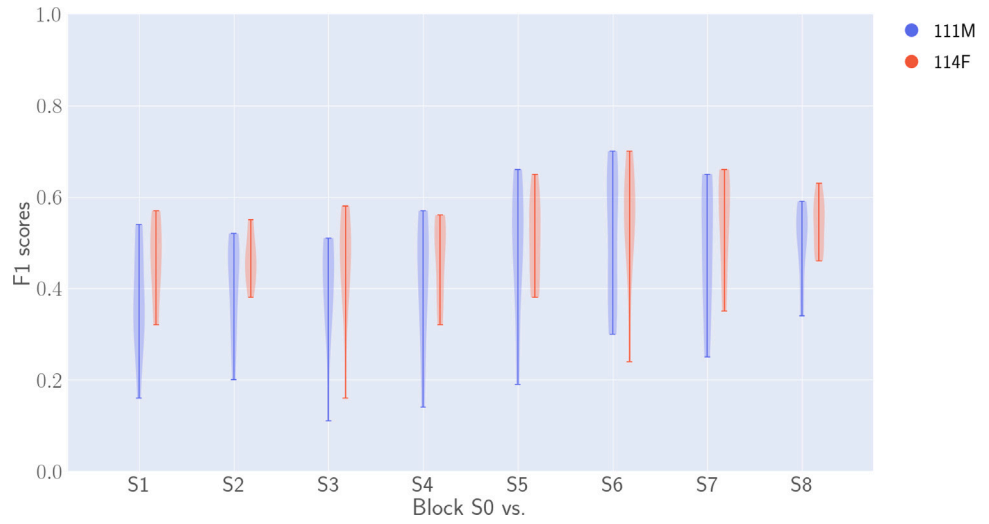


Fig. 8. Speaker-wise (111M, 114F) distribution of F1 scores for each block derived from the speaker-independent personalised experiments.

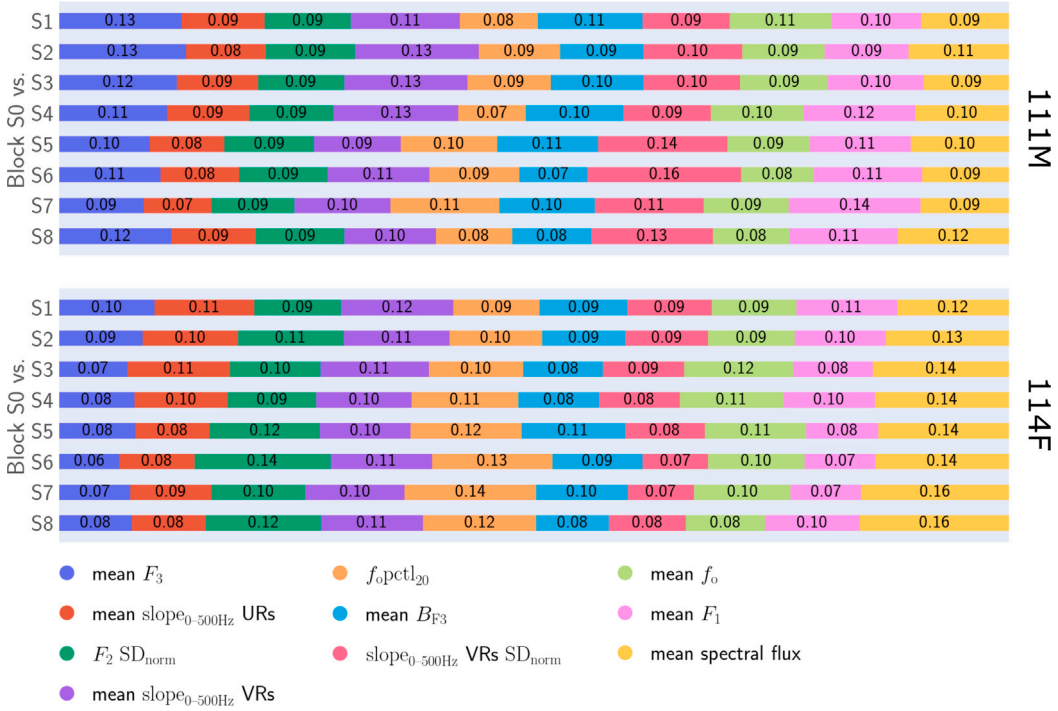


Fig. 9. Speaker-wise (111M, 114F) random forest feature importances derived from the speaker-independent personalised experiments. B_{F3} = bandwidth of third vowel formant, f_0 = fundamental frequency, $F_{1|2|3}$ = frequency of first|second|third vowel formant, pctl$_{20}$ = 20th percentile, SD$_{norm}$ = standard deviation normalised by the arithmetic mean (coefficient of variation), URs = unvoiced regions, VRs = voiced regions.

4.1. Limitations

Even though our study provides novel insights into the acoustic manifestation of potential vocal fatigue paving the way for an automatic assessment of long-term variations of the human voice, our findings have to be interpreted carefully. On the one hand, this is due to our pilot dataset consisting of just six native German speakers. On the other hand, we decided to keep the cohort for our basic analyses as consistent as possible and included male speakers in our *VocDoc-pilotSet* only. Moreover, one of the four speakers of the *VocDoc-pilotSet* as well as the male speaker of the *VocDoc-pilotSet extension* had a significantly higher age than the other speakers. Finally, we want to point out that, reading out loud a text from a book does not represent an authentic everyday life setting. A larger as well as gender- and age-balanced dataset recorded 'in the wild' is thus warranted for further investigations, also allowing for the utilisation of more sophisticated machine learning techniques, such as deep neural networks. The influence of language on the manifestation of vocal fatigue should be studied as well.

4.2. VocDoc app: Status quo and future recommendations

Promoting a mobile solution for objectively capturing clinically relevant long-term variations of the voice in everyday life settings, we developed a smartphone application prototype – the *VocDoc* – in collaboration with the audEERING GmbH (https://audeering.com [as of 29 March 2023]), which shall be further developed on the basis of findings of this study. The *VocDoc* is expected to be utilised together

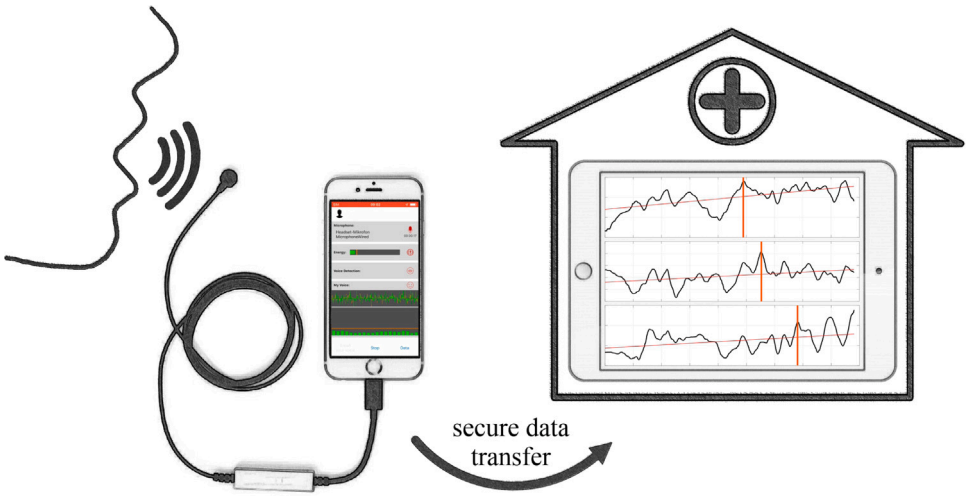


Fig. 10. Basic concept of robust and privacy-compliant long-term voice recording and vocal feature extraction in the patient's natural environment by means of the *VocDoc* application and subsequent data analysis at the hospital for diagnostic support and intervention planning.

with a clip-on microphone placed at the collar. The current version of the *VocDoc* initially asks the user to speak a few sentences in a silent environment. Thereby, intrinsic voice parameters of the specific user are retrieved for the optimisation of an intelligent noise-robust pre-trained voice activity detector. Subsequently, the prototype is ready for use in real-world settings and manages to reliably capture the voice of the user within potential environments of more than one speakers present. For data privacy reasons, the *VocDoc* immediately extracts a set of acoustic features from each detected voice segment, which is stored in the form of a vector together with a time stamp in a local text file. Raw audio data are not stored. The *VocDoc* can be used until the user manually stops recording or until the smartphone runs out of battery. Finally, the text file with the stored acoustic features can be exported and analysed. In the future, we envisage patients with (subjective) voice problems to be equipped with the *VocDoc*. Over an initial period of a few days or weeks, voice data shall be collected in order to identify a speaker-specific set of optimal acoustic features. Thenceforth, the *VocDoc* can be used in everyday settings to collect representative long-term voice data of the patient. Whenever a recording is done, the respective feature text file is transmitted via a secure channel to the corresponding healthcare facility for data interpretation by the attending doctor or SLT. The general concept of our *VocDoc* is illustrated in Fig. 10. The *VocDoc* is intended to support and facilitate both future diagnostic and intervention processes in phoniatric patients. Moreover, the same framework of automatically capturing acoustic long-term voice information by means of a smartphone could also be used for an (earlier) identification of vocal changes that originate from potential respiratory, neurological, or mental conditions.

In an immediate next step, voice data from an age- and gender-balanced cohort of patients affected by (subjective) voice problems shall be collected 'in the wild' by means of the *VocDoc*, and analysed for pathological changes over time. Thereby, both voice disorder-specific information that cannot be retrieved at the clinical site in the framework of a standard voice examination shall be gained for the very first time, and the usability of the *VocDoc* as well as the patients' willingness of its use in everyday settings shall be evaluated.

5. Conclusion

In this study, we explored physiological vocal fatigue from three different perspectives, namely the speaker's subjective perspective, the perspective of SLTs, and the engineering perspective with respect to an automatic recognition of elapsed speaking time. Even though we found a small set of acoustic candidate features to universally describe vocal change over time, i. e., the mean spectral slopes from 0–500 Hz in voiced and unvoiced regions and the mean bandwidth of the third vowel formant, our results reveal that vocal fatigue has individual effects across different speakers. Nevertheless, we could demonstrate that vocal changes occur – even in voice healthy speakers – and that machine learning methodology has the potential to automatically detect and characterise them when being trained on data from the same speaker. Knowledge gained in this study shall contribute to the (further) development of a mobile application to promote a clinical delineation of pathological long-term voice variations in everyday life settings.

CRediT authorship contribution statement

Florian B. Pokorny: Designed the study, Coordinated the implementation of the study, Prepared and carried out the listening experiments with the SLTs, Analysed parts of the data, Wrote the Introduction, Discussion, and Conclusion sections of the first draft of the manuscript, Writing – review & editing. **Julian Linke:** Collected and prepared the study dataset, Analysed parts of the data, Supervised the machine learning experiments, Wrote the Materials and methods as well as the Results sections of the first draft of the manuscript, Writing – review & editing. **Nico Seddiki:** Collected and prepared the study dataset, Conducted the machine learning experiments, Helped with the first draft of the Materials and methods as well as the Results sections, Writing – review & editing. **Simon Lohrmann:** Collected and prepared the study dataset, Assisted in the data analysis, Writing – review & editing. **Claus Gerstenberger:** Collected and prepared the study dataset, Provided consultancy in biomedical engineering and phoniatrics matters, Writing – review & editing. **Katja Haspl:** Facilitated the interpretation of data and findings from an SLT's perspective, Writing – review & editing. **Marlies Feiner:** Facilitated the interpretation of data and findings from an SLT's perspective, Writing – review & editing. **Florian Eyben:** Consulted in matters of acoustic feature analysis and feature interpretation, Writing – review & editing. **Martin Hagmüller:** Provided methodological advice and helped with the overall interpretation of the results, Writing – review & editing. **Barbara Schuppler:** Provided methodological advice and helped with the overall interpretation of the results, Writing – review & editing. **Gernot Kubin:** Supervised the technical design and implementation, Writing – review & editing. **Markus Gugatschka:** Conceptualised and supervised the overall implementation of the study, Writing – review & editing.

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.

Data availability

Data will be made available on request.

Acknowledgements

The authors want to express their gratitude to the speakers of the *VocDoc-pilotSet extension* for providing us recordings of their lectures for the purpose of vocal fatigue analysis in the framework of this study.

References

[1] K. Verdolini, L.O. Ramig, Occupational risks for voice problems, Logop. Phoniatr. Vocol. 26 (1) (2001) 37–46.
[2] S.M. Cohen, W.D. Dupont, M.S. Courey, Quality-of-life impact of non-neoplastic voice disorders: A meta-analysis, Ann. Otol. Rhinol. Laryngol. 115 (2) (2006) 128–134.
[3] S.M. Cohen, J. Kim, N. Roy, C. Asche, M. Courey, The impact of laryngeal disorders on work-related dysfunction, Laryngoscope 122 (7) (2012) 1589–1594.
[4] P.H. Dejonckere, P. Bradley, P. Clemente, G. Cornut, L. Crevier-Buchman, G. Friedrich, P. Van De Heyning, M. Remacle, V. Woisard, A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques, Eur. Arch. Otorhinolaryngol. 258 (2) (2001) 77–82.
[5] G. Friedrich, P.H. Dejonckere, The voice evaluation protocol of the European laryngological society (ELS) – first results of a multicenter study, Laryngo-Rhino-Otologie 84 (10) (2005) 744–752.
[6] G. Friedrich, Basisprotokoll für die stimmdiagnostik – richtlinien der European laryngological society (ELS), Forum Logopädie 20 (4) (2006) 6–12.
[7] B. Schneider-Stickler, W. Bigenzahn, Stimmdiagnostik: Ein Leitfaden Für Die Praxis, Springer, 2013.
[8] A. Panesar, Machine Learning and AI for Healthcare, Springer, 2019.
[9] D. Dias, J. Paulo Silva Cunha, Wearable health devices – vital sign monitoring, systems and technologies, Sensors 18 (8) (2018) 2414.
[10] M. Sheikh, M. Qassem, P.A. Kyriacou, Wearable, environmental, and smartphone-based passive sensing for mental health monitoring, Front. Digit. Health 3 (2021) 662811.
[11] S. Liu, J. Han, E.L. Puyal, S. Kontaxis, S. Sun, P. Locatelli, J. Dineley, F.B. Pokorny, G. Dalla Costa, L. Leocani, A.I. Guerrero, C. Nos, A. Zabalza, P.S. Sø rensen, M. Buron, M. Magyari, Y. Ranjan, Z. Rashid, P. Conde, C. Stewart, F.A. A, R.J.B. Dobson, R. Bailón, S. Vairavan, N. Cummins, V.A. Narayan, M. Hotopf, G. Comi, B. Schuller, RADAR-CNS Consortium, Fitbeat: COVID-19 estimation based on wristband heart rate using a contrastive convolutional auto-encoder, Pattern Recognit. 123 (2022) 108403.
[12] M.P. Gelfer, M.L. Andrews, C.P. Schmidt, Effects of prolonged loud reading on selected measures of vocal function in trained and untrained singers, J. Voice 5 (2) (1991) 158–167.
[13] J.C. Stemple, J. Stanley, L. Lee, Objective measures of voice production in normal subjects following prolonged voice use, J. Voice 9 (2) (1995) 127–133.
[14] L. Rantala, P. Lindholm, E. Vilkman, F0 change due to voice loading under laboratory and field conditions. A pilot study, Logop. Phoniatr. Vocol. 23 (4) (1998) 164–168.
[15] L. Rantala, L. Paavola, P. Körkkö, E. Vilkman, Working-day effects on the spectral characteristics of teaching voice, Folia Phoniatr. Logop. 50 (4) (1998) 205–211.
[16] L. Rantala, E. Vilkman, Relationship between subjective voice complaints and acoustic parameters in female teachers' voices, J. Voice 13 (4) (1999) 484–495.
[17] E. Vilkman, E.-R. Lauri, P. Alku, E. Sala, M. Sihvo, Effects of prolonged oral reading on F0, SPL, subglottal pressure and amplitude characteristics of glottal flow waveforms, J. Voice 13 (2) (1999) 303–312.
[18] R. Arya, S. Bagwan, S. Relekar, Vocal fatigue in school teachers and it's relation to the acoustic analysis of voice, Indian J. Otolaryngol. Head Neck Surg. 74 (2) (2022) 1979–1988.
[19] S.P. Bayerl, D. Wagner, I. Baumann, T. Bocklet, K. Riedhammer, Detecting vocal fatigue with neural embeddings, J. Voice (2023) http://dx.doi.org/10.1016/j.jvoice.2023.01.012.
[20] B. Pfister, T. Kaufmann, Sprachverarbeitung - Grundlagen und Methoden der Sprachsynthese und Spracherkennung, 2. Auflage, Spinger Vieweg, Germany, 2017.
[21] F. Eyben, M. Wöllmer, B. Schuller, openSMILE: The munich versatile and fast open-source audio feature extractor, in: Proceedings of the ACM International Conference on Multimedia, ACM, Florence, Italy, 2010, pp. 1459–1462.
[22] F. Eyben, F. Weninger, F. Groß, B. Schuller, Recent developments in openSMILE, the Munich open-source multimedia feature extractor, in: Proceedings of the ACM International Conference on Multimedia, ACM, Barcelona, Spain, 2013, pp. 835–838.
[23] F. Eyben, F. Weninger, S. Squartini, B. Schuller, Real-life voice activity detection with LSTM recurrent neural networks and an application to hollywood movies, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Vancouver, Canada, 2013, pp. 483–487.
[24] F. Eyben, K.R. Scherer, B.W. Schuller, J. Sundberg, E. André, C. Busso, L.Y. Devillers, J. Epps, P. Laukka, S.S. Narayanan, K.P. Truong, The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing, IEEE Trans. Affect. Comput. 7 (2) (2016) 190–202.
[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[26] P.A. Keating, M. Garellek, J. Kreiman, Acoustic properties of different kinds of creaky voice, in: Proceedings of the International Congress of Phonetic Sciences, IPA, Glasgow, Scotland, 2015, pp. 2–7.
[27] S.A. Memon, Acoustic correlates of the voice qualifiers: A survey, 2020, http://dx.doi.org/10.48550/arXiv.2010.15869, arXiv preprint.
[28] B. Tracey, D. Volfson, J. Glass, M. Kostrzebski, J. Adams, T. Kangarloo, A. Brodtmann, E. Dorsey, A. Vogel, Towards interpretable speech biomarkers: Explaining MFCC2, Res. Square Preprint (2023) http://dx.doi.org/10.21203/rs.3.rs-2478289/v1.
[29] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust. Speech Signal Process. 28 (4) (1980) 357–366.
[30] C.O. Dumitru, I. Gavat, A comparative study of feature extraction methods applied to continuous speech recognition in Romanian language, in: Proceedings of the International Symposium on Electronics in Marine, IEEE, 2006, pp. 115–118.
[31] C. Ittichaichareon, S. Suksri, T. Yingthawornsuk, Speech recognition using MFCC, in: Proceedings of the International Conference on Computer Graphics, Simulation and Modeling, Vol. 9, 2012, pp. 135–138.
[32] A.S. Haq, M. Nasrun, C. Setianingsih, M.A. Murti, Speech recognition implementation using MFCC and DTW algorithm for home automation, Proc. Electr. Eng. Comput. Sci. Inform. 7 (2) (2020) 78–85.
[33] V. Tiwari, MFCC and its applications in speaker recognition, Int. J. Emerg. Technol. 1 (1) (2010) 19–22.
[34] U. Ayvaz, H. Gürüler, F. Khan, N. Ahmed, T. Whangbo, A. Bobomirzaevich, Automatic speaker recognition using mel-frequency cepstral coefficients through machine learning, Comput. Mater. Contin. 71 (3) (2022).
[35] H. Dolka, A.X. VM, S. Juliet, Speech emotion recognition using ANN on MFCC features, in: Proceedings of the International Conference on Signal Processing and Communication, IEEE, 2021, pp. 431–435.
[36] L. Juvela, B. Bollepalli, X. Wang, H. Kameoka, M. Airaksinen, J. Yamagishi, P. Alku, Speech waveform synthesis from MFCC sequences with generative adversarial networks, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2018, pp. 5679–5683.
[37] L.E. Boucheron, P.L. De Leon, S. Sandoval, Low bit-rate speech coding through quantization of mel-frequency cepstral coefficients, IEEE Trans. Audio Speech Lang. Process. 20 (2) (2011) 610–619.
[38] K.D. Bartl-Pokorny, F.B. Pokorny, A. Batliner, S. Amiriparian, A. Semertzidou, F. Eyben, E. Kramer, F. Schmidt, R. Schönweiler, M. Wehler, et al., The voice of COVID-19: Acoustic correlates of infection in sustained vowels, J. Acoust. Soc. Am. 149 (6) (2021) 4377–4383.
[39] M. Malik, M.K. Malik, K. Mehmood, I. Makhdoom, Automatic speech recognition: A survey, Multimedia Tools Appl. 80 (2021) 9411–9457.