Biomedical Signal Processing and Control 86 (2023) 105201

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Automated voice pathology discrimination from audio recordings benefits from phonetic analysis of continuous speech

Mark Huckvale [*], Zhuoya Liu [1], Catinca Buciuleac

Speech, Hearing and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, United Kingdom

ARTICLE INFO

Keywords:
Voice disorders
Continuous speech
Machine learning
Diagnosis
Health applications

ABSTRACT

In this paper we evaluate the hypothesis that automated methods for diagnosis of voice disorders from speech recordings would benefit from contextual information found in continuous speech. Rather than basing a diagnosis on how disorders affect the average acoustic properties of the speech signal, the idea is to exploit the possibility that different disorders will cause different acoustic changes within different phonetic contexts. Any differences in the pattern of effects across contexts would then provide additional information for discrimination of pathologies. We evaluate this approach using two complementary studies: the first uses a short phrase which is automatically annotated using a phonetic transcription, the second uses a long reading passage which is automatically annotated from text. The first study uses a single sentence recorded from 597 speakers in the Saarbrucken Voice Database to discriminate structural from neurogenic disorders. The results show that discrimination performance for these broad pathology classes improves from 59% to 67% unweighted average recall when classifiers are trained for each phone-label and the results fused. Although the phonetic contexts improved discrimination, the overall sensitivity and specificity of the method seems insufficient for clinical application. We hypothesise that this is because of the limited contexts in the speech audio and the heterogeneous nature of the disorders. In the second study we address these issues by processing recordings of a long reading passage obtained from clinical recordings of 60 speakers with either Spasmodic Dysphonia or Vocal fold Paralysis. We show that discrimination performance increases from 80% to 87% unweighted average recall if classifiers are trained for each phone-labelled region and predictions fused. We also show that the sensitivity and specificity of a diagnostic test with this performance is similar to other diagnostic procedures in clinical use. In conclusion, the studies confirm that the exploitation of contextual differences in the way disorders affect speech improves automated diagnostic performance, and that automated methods for phonetic annotation of reading passages are robust enough to extract useful diagnostic information.

1. Introduction

The human voice production system can become impaired in multiple ways involving structural, neurogenic, inflammatory, or muscle tension disorders [1]. A challenge for the clinician is discriminating between disorders, since they may have very different consequences for therapy. In particular, cancers of the throat and larynx require urgent detection and treatment. Instrumental methods for endoscopic examination of the larynx and objective assessments involving acoustic and Electroglottographic (EGG) analysis are available, but only in specialized centres, and typically only after screening by non-instrumental means. Differentiation between types of disorders by subjective auditory assessments alone is difficult because of similarities in the auditory effect of different pathologies, and diagnostic reliability is highly influenced by clinician training, background, and experience [2]. Recently, machine learning approaches for objective assessment of voice pathology have become popular, since they hold the promise of accurate pathology detection and discrimination from simple audio recordings, while making screening for pathology more accessible [3,4,5]. Although there are many such studies focusing on contrasting pathological from non-pathological voices (see [6] for a survey), only a few studies in the past decade have investigated differential diagnosis of voice pathologies [7,8,9]. This paper is concerned with the assessment and improvement of the best current machine learning methods for voice pathology

[*] Corresponding author.
E-mail addresses: m.huckvale@ucl.ac.uk (M. Huckvale), zhuoya.liu.20@alumni.ucl.ac.uk (Z. Liu), catinca.buciuleac.16@alumni.ucl.ac.uk (C. Buciuleac).
[1] Present address: Research Centre for Language, Cognition, and Neuroscience, The Hong Kong Polytechnic University.

https://doi.org/10.1016/j.bspc.2023.105201
Received 27 October 2022; Received in revised form 13 June 2023; Accepted 21 June 2023
Available online 5 July 2023

discrimination.

The most common approach to automated voice disorder assessment has been to analyse sustained vowel productions of the speaker (e.g., /ɑ/ , /e/, and /i/) instead of continuously spoken speech. Although sustained vowels have long been used by clinicians to assess voice, they are clearly unrepresentative of everyday speech [10] and do not fully exercise the voice production mechanism. Since continuous speech requires the exercise of more laryngeal functions, it seems likely that this style would better expose voice disorders. There has also been some recent evidence that continuous speech recordings can allow for better automated pathology detection than sustained vowel recordings [11,12,13]. However, automated assessment of continuous speech can be challenging for machine learning methods because of the increase in acoustic variability caused by the verbal content of the speech. While an isolated vowel can be said to have relatively stationary spectral properties and thus can be characterized by averages made over a whole recording, the same cannot be said of a read passage. It seems likely that different phonetic elements in speech will put different stresses on a disordered voice, since the production of different phones is associated with different vocal tract and laryngeal configurations. The consequence is that an average of signal properties computed over all phones in a recording could dilute variation that is useful for identification or discrimination of pathologies.

Examples of interactions between pathology and phonetic context include the starting and stopping of voicing in plosives, which might be affected by disorders of adduction and abduction, or the aerodynamic interactions between vocal tract constriction and voicing in voiced fricatives, which might expose disorders affected by reduced glottal air flow. A previous study has shown how in dysphonic voice with vocal fold thickening, unstressed syllables are more likely to be produced with insufficient subglottal pressure realizing aphonia [14]. Even in normal voices, variation of voice quality with phonetic context has been found in studies such as [15,16] and predicted by phoneticians [17]. A few studies have also looked at variation in voice quality with phonetic context for the assessment of Parkinson's [18] and on the assessment of severity of voice disorder [19]. These studies have exploited contexts such as manner of articulation (e.g., plosives, fricatives, and affricates), voicing (e.g., voiced and voiceless onsets), and the height of the tongue (e.g., high vowels and low vowels) to improve assessments.

Not only might voice disorders have greater consequences in certain phonetic contexts, but there might be an interaction between the type of disorder and the size and nature of its effects in those contexts. Where one disorder might have one acoustic consequence in a particular context, another disorder might have a different effect. Thus, an analysis of changes in voice quality with phonetic context might provide useful information for the discrimination of voice pathologies. The goal of this paper is to test the utility of the information arising from this interaction of pathology and context using automated voice pathology discrimination from phonetically annotated continuous speech recordings.

This paper describes two complementary studies into the use of automated methods for the diagnosis of voice disorder from audio recordings that test the phonetic context interaction hypothesis. In the first study, we explore the use of a state-of-the-art method for voice pathology classification using a standard database of normal and pathological speech. The goal is to test whether discrimination of two broad pathology classes from an audio recording of continuous speech is better when the speech is treated as a collection of different phones rather than as a whole. Using automated phonetic transcription alignment, we show that phonetic context information is useful, but that performance of the method may be constrained by limitations of the database which comprises numerous distinct pathologies and only a single, short instance of continuous speech for each speaker. Therefore, in the second study, we extend the method by applying it to a longer read passage and to two narrowly specified disorders. Again, we explicitly compare discrimination performance using the passage as a whole with performance when the method is applied independently to different phone-labelled regions in the passage and the predictions fused. To establish the phonetic contexts in the reading passage we introduce a method based on automated alignment from text rather than from phonetic transcription and evaluate its performance. The outcome of the second study not only confirms that interactions between pathology and phonetic context provide useful information for diagnosis, but that the automated methods are robust enough to deliver diagnostic performance similar to tests currently used in clinical practice. In a final section we reflect on limitations of the current study and make proposals for future work.

2. Study 1. Evaluation of automated voice pathology discrimination using a common database and a short phrase

There have been many published studies that use machine learning approaches for the automated detection of voice disorders from audio recordings using isolated vowels (see [6] for review), but fewer studies have investigated discrimination between disorders, and few of those have analysed continuous speech. The goal of this study is to evaluate a method for exploiting the phonetic contexts in continuous speech for discriminating between voice pathologies. We implement a state-of-the-art system using short phrase recordings from the Saarbrücken Voice Database (SVD) [20] to differentiate between two broad classes of disorders. We compare the classification performance of the system when the phrases are analysed as a whole with performance when the phrases are treated as a collection of different phonetic contexts.

2.1. Background to the corpus

The Saarbrücken Voice Database contains material from a total of 2225 recording sessions from both sufferers of various voice disorders (454 male and 548 female speakers) and healthy control speakers (423 male and 428 female). The age of speakers varies from 6–94 years (pathological) and 9–84 years (control), and there are an average of 1.2 recording sessions per speaker (max = 24). Each recording session represents a set of audio and electroglottographic recordings for both a short phrase (i.e. "Guten Morgen,wie geht es Ihnen?" ("Good morning, how are you?") and sustained vowels on various pitches (i.e. /i/, /a/ and /u/, in isolation and together, on typical, higher, lower, rising and falling pitch). Both the audio and EGG recordings are available sampled at 16-bit precision at 50 k samples/sec.

The database uses 71 different pathology labels for its recordings, though 263 sessions are assigned more than one label [21]. Some pathologies are much better represented than others – the top three most frequently occurring pathologies (i.e. vocal fold paralysis, hyperfunctional dysphonia, laryngitis) each have more than 80 associated recordings, while there are 19 pathologies which only occur once.

While the SVD is an extremely useful resource, it is not an easy database to partition for use in machine learning. The imbalance in the frequency of pathologies, the assignment of multiple pathologies to recordings, and the presence of multiple recordings per speaker could easily bias classification performance. Also, any cross-validation process that did not take speaker into account could allow the same speaker to be present in both training and testing partitions, artificially boosting performance. Our evaluation will use a standardized way of working with the multiple recordings and multiple diagnoses per speaker found in the SVD as described in [5].

2.2. Database subset selection

We explore two major classes of pathologies available in the SVD: those that arise from structural disorder in the larynx itself ('structural'), and those that arise from disorder of the nervous control of the larynx ('neurogenic'). Structural disorders include laryngitis, nodules and polyps; neurogenic disorders include vocal fold paresis and spasmodic dysphonia. Table 1 provides the diagnostic labels used in the SVD to define the two groups.

Table 1
Summary of pathology labels used in the SVD to define the two pathology subsets.

Structural	Neurogenic
Kontaktpachydermie (Pachydermia)	Rekurrensparese (Recurrent laryngeal nerve paresis/ Vocal fold paresis)
Laryngitis	Spasmodische Dysphonie (Spasmodic dysphonia)
Leukoplakie (Leukoplakia)	
Reinke Ödem (Reinke's Edema)	
Stimmlippenkarzinom (Vocal fold carcinoma)	
Stimmlippenpolyp (Vocal fold polyps)	

As defined, there are 325 recordings with structural disorders and 272 recordings with neurogenic disorders in the SVD.

2.3. Audio selection and pre-processing

For these experiments, we used the recording of the spoken phrase which we have transcribed as / g u t ə n m ɔ g ə n v i g e t e s i n ə n /. Alignment of the transcription to the audio was performed using the 'analign' tool in SFS [22]. This uses a pre-trained set of HMM phone models based on MFCC spectral analyses of the signal. Audio materials were downsampled to 16 kHz for the feature extraction used for classification.

2.4. Feature extraction and normalisation

The OpenSMILE toolkit [23] was used to extract features using the ComParE feature set [24] as used in the 2013 Interspeech Computational Paralinguistics challenge. This feature extraction system first computes 126 low-level descriptors (LLDs) from the signal at 100 feature frames per second. These features describe short-term spectral and temporal characteristics of the signal and other extracted measures relating to voice quality. These features may then be summarised over the recording using a set of functionals which provide averages and dispersion measures of various kinds. In this study we either use the full set of functionals found in the ComParE13 configuration, which delivers 6373 features, or we simply compute a median and an inter-quartile range for each LLD over each region of interest, delivering 252 features. Individual features were normalized by brute force Gaussianisation using the bestNormalize package [25] in 'R', which maps the rank of each value into a sample from a cumulative Gaussian pdf. Normalization was performed as part of cross-validation, such that only the training data in each fold were used to define the normalizing transform.

2.5. Classifier construction

The SVM classifier was implemented in R using the e1071 package [26]. A radial basis function kernel was used, and a simple grid-search established to find optimum values for the cost parameter, which was then fixed for all classifiers. The gamma parameter was set to the reciprocal of the number of features. The SVM classifier was run in a mode that delivers the classification probability rather than the best class. This allows the post-hoc calculation of a Receiver Operating Characteristic curve (ROC) and the determination of the Area-Under-Curve (AUC) metric of performance for each test condition. Using the optimal threshold from the ROC, we are able to report the best Unweighted Average Recall (UAR) for each condition. The UAR is an estimate of the accuracy of the system if the test data had equal numbers of each pathological class.

The SVM was trained and evaluated using five-fold cross-validation. Assignment of recordings to the cross-validation fold was done on the basis of speaker number, to ensure that the same speaker did not appear in more than one fold.

2.6. Experimental conditions

Three experimental conditions were evaluated to test the hypothesis. The first baseline condition mirrors evaluations performed in our previous study [5] on pathology detection using the SVD, but now using the two pathology subsets. For this baseline, all the 6373 SMILE features were used, computed over each whole recording. A second baseline used only the 252 features computed from the LLDs for the whole phrase. The test condition trains a separate SVM for each of the 11 phone types found in the annotated phrase. This takes as input the mean and IQR of the LLDs for frames labelled with that phone. The eleven phone SVMs then deliver a classification probability per recording, and these are fused using linear discriminant analysis (LDA) to generate one overall discriminant between the two classes. The discriminants were computed using cross-validation such that the phone probabilities for the sample under test were not used in the calculation of the LDA weights. The discriminant was then used to create an ROC curve to compare with the whole-phrase classifiers.

2.7. Experimental results and discussion

Classification performance in terms of AUC and UAR for the three conditions are shown in Table 2.

In discussing the outcomes, it should be noted that an AUC of 0.5 and a UAR of 50% represent random classification. Results show that the full OpenSMILE feature set computed from a large number of functionals applied to the LLDs does not provide any advantage over a simpler feature set using median and IQR. This is probably because the shortness of the phrase makes the complex functional averages unreliable, effectively adding noise to their estimates. The analysis by individual phone types in the phrase followed by score fusion with LDA is superior to both baseline conditions which compute an average over the whole phrase. This is probably because some elements of the phrase are more sensitive to differences in pathology than others, and the LDA process can emphasise the more important elements and downplay others. A post-hoc analysis of independently trained phone classifiers shows that the best classification accuracy arises from phones /g, ə, e/, while poorer accuracy is obtained from regions labelled with /t, i, v/. Of course, the specific lexical, syllabic and prosodic context for these phones will also play a part in their utility for classification.

To understand the possibility of using such a method in a clinical application for diagnosing voice pathologies, it is instructive to plot the value of the discriminant based on the true pathological class of the speakers, and a Receiver Operating Characteristic (ROC) showing the trade-off between sensitivity and specificity, see Fig. 1. To achieve a diagnostic test with a sensitivity of 0.9 for structural disorders, the method would deliver a specificity of only 0.4, that is, to detect 90% of all structural disorders, the cost would be the misidentification of 60% of neurogenic disorders.

In summary, the contextualised phonetic analysis of the short phrase in the SVD does lead to a small increase in the ability to discriminate structural from neurogenic disorders. This supports the hypothesis that different disorders will have different effects in different phonetic

Table 2
Classification performance of three test conditions for pathology discrimination from the short phrase.

Condition	# Features	AUC	UAR%
SMILE full, whole phrase	6373	0.584	59.41
LLD, whole phrase	252	0.661	63.21
LLD, per phone, LDA fusion	252	0.721	67.30

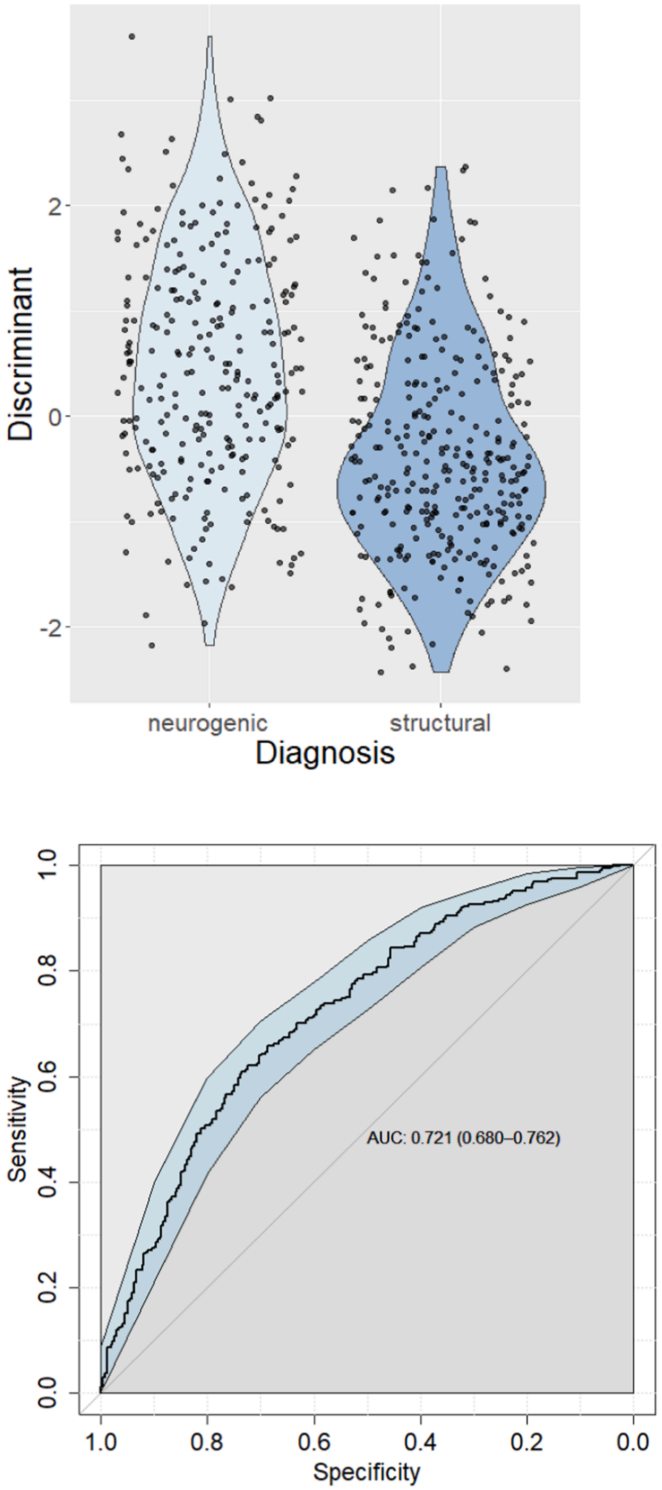

Fig. 1. Distribution of the phone LDA discriminant for the true pathology classes and an ROC showing tradeoff between sensitivity and specificity for different thresholds of the discriminant.

contexts, and that sensitivity to context improves discrimination. However, the overall performance is still rather low, and perhaps too low to be useful in a clinical application. If the phonetic context hypothesis is correct, greater discrimination should come from longer spoken passages which will have more phonetic contexts and more speech over which to compute averages. The use of the broad classes 'structural' and 'neurogenic' may also be a hindrance in that these classes may hide a heterogeneous collection of acoustic effects in each context. Thus, a good direction for other studies would be to narrow down to specific disorders, which might have a more consistent set of acoustic effects. We consider both of these improvements in the next study.

3. Study 2 voice pathology discrimination from analysis of a read passage

This study advances the method described in the study above in two ways: it is based on recordings of a long reading passage rather than a short phrase, and it evaluates a classifier that discriminates between two specific voice pathologies. The goal is to investigate whether contextualised phonetic analysis can also provide performance improvements when used with a greater length of speech material and more narrowly-defined disorders. An innovative aspect of the work is the use of an automated phonetic annotation technique to segment a long recording using only a text transcription. A further outcome of the study is a post-hoc analysis of which phonetic contexts are more useful in detecting specific pathologies, which might lead to the design of specific speech materials for diagnosis. An earlier version of this study was previously published as [31].

3.1. Source of data

The study used previously collected 'Arthur the Rat' passage reading and sustained vowel production recordings made from individuals (British English speakers) presenting at a specialist multidisciplinary voice clinic. There are 38 participants subsequently diagnosed with Spasmodic Dysphonia (SD) (6 Abductor, 32 Adductor) and 22 participants diagnosed with Vocal fold Palsy (VP, otherwise known as vocal fold paralysis). The mean age for SD speakers (10 male, 28 female) was 62 ± 15 years. The mean age for VP speakers (20 male, 2 female) was 53 ± 22 years. The choice of these two pathologies was due to data availability, but they do reflect two disorders with different aetiologies and therapies.

SD and VP are two distinct types of neurogenic voice disorder. SD is a form of focal dystonia. There are two main phenotypes, both characterized by abrupt spasms of intrinsic laryngeal muscles. The commoner form, Adductor SD (90%), is associated with spasmodic closure of the vocal folds (i.e., glottal stopping) particularly following voiced onsets. This results in involuntary phonatory breaks during propositional speech and in addition the voice has a strained/ strangled quality. The less common form, Abductor SD (10%), is associated with involuntary spasmodic opening of the vocal folds (i.e., glottal widening). It is also associated with unnatural breathy or aphonic interludes during phonation, and is worsened by the use of voiceless consonants, prolonging word or sentence duration. In both forms, speech becomes slower, more effortful, and more dysfluent with increasing severity, but less affected during whispering and non-speech vocalizations, such as laughter and crying. VP occurs when there is neural damage to the intrinsic muscles of the larynx due to viral neuropathy, neck or thoracic surgery, cancer, neck trauma or other neurologic conditions. People with VP may have a hoarse, weak, breathy, or diplophonic voice, with loss of volume and elevation in pitch [27].

Speech and EGG recordings were made with Laryngograph hardware, which used an electret microphone placed on the EGG neckband, in a quiet clinic room. Most recordings were made at 44,100 samples/sec 16-bit, while some were at 22,050 samples/sec. Only the recorded speech signals were used in this study, and the EGG recordings will be analysed in a later study.

3.2. Audio pre-processing

Each speaker produced some isolated words and a reading passage. The passage was a recording of the 'Arthur the Rat' story, which was on average 149 s of audio for the SD speakers and 141 s for the VP speakers.

For the baselines, two types of vowel-sound extracts were segmented from recordings of the production of sustained vowels collected in another assessment:

- IY: instance of an /i/ vowel spoken on a low pitch.
- AE: instance of an /æ/ vowel in the isolated word "sat".

For the passage reading, manual editing of the audio was required to eliminate any speech from the clinician prompting the speaker before or after each extract. However, to maintain consistency, any clinician's speech that overlapped with the participant's speech was retained. All signals were then resampled to 16,000 samples/sec.

3.3. Alignment and annotation

An edited transcript of the reading passage was created separately for each speaker to make transcriptions that matched the actual production. There were 3 out of 60 transcripts that needed major edits due to the deletion of whole sentences. For the rest of the recordings, only a few manual corrections were required when the participants repeated or changed occasional words. The orthographic transcript was then aligned with the speech audio using the Montreal forced aligner [28]. This forced alignment approach produced a segmentation of the signal at both word and phone levels. The phonetic annotation was based on an American English pronunciation dictionary with 41 phone types. The alignment and phonetic labelling permitted the analysis of phonetic context within the pathological speech recording, as the individual acoustic segments corresponding to individual phones could be grouped together for voice disorder assessment. Fig. 2 shows the examples of the automatic alignment and annotation of the word 'Arthur'.

To estimate the accuracy of the automated alignment, a random sample of 20 recordings was chosen, and then a random section of 3 s of each was selected for manual checking. Of 493 annotations checked, only 9 (2%) were found to be in significant error (greater than 10 ms from a satisfactory ideal position). On average, those in error were shifted by 30 ms from their preferred location. Based on these results, no corrections to the automatic phonetic annotations were made for the experiment. While the alignment seems good, it is possible that alignment might have been improved had a British English dictionary been used.

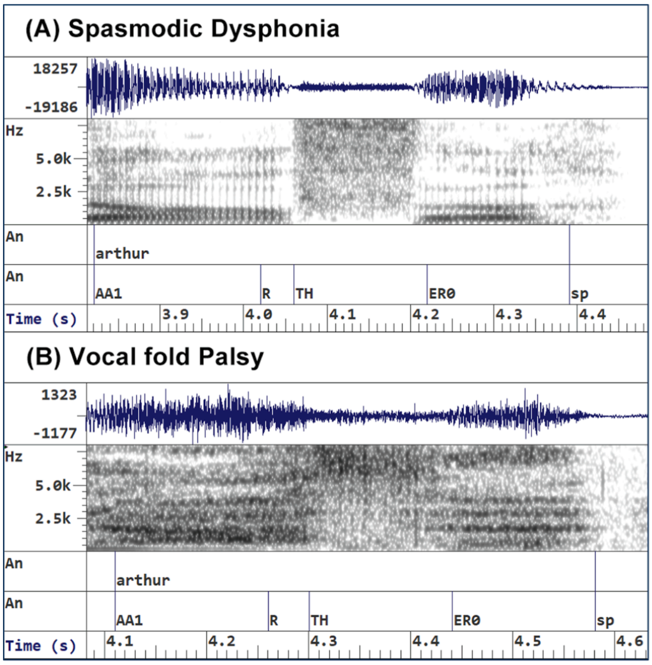

Fig. 2. Examples of automatic alignment and annotation. Voice recordings are from a speaker with SD (above) and a speaker with VP (below) producing the word 'Arthur'.

3.4. Feature extraction and normalisation

The OpenSMILE analysis system was used to extract features for processing as in Study 1. Two strategies for summarizing features across recordings, as before. In the Functional strategy, we used the large set of summary functionals found in the COMPARE13 configuration of OpenSMILE [29], which delivered 6373 features per recording. In the Summary strategy, we used the COMPARE13 low-level descriptors (LLD) configuration delivering 126 features per 10 ms frame and then computed the median and inter-quartile range of each LLD to give 252 features for each specified region of a recording. This latter approach allowed us to generate a feature vector that summarised the features found in specified phonetic contexts identified by the phone labels on each frame. All features were normalised using z-scores. For the Functional strategy, feature selection was performed on the basis of an F-ratio statistic to select the 1000 most active features for discrimination. For the Summary strategy, feature selection was not conducted.

3.5. Classifier construction

An SVM classifier was used as before, using a radial basis function kernel and a cost parameter C = 2 and gamma set to the reciprocal of the number of features. For classification, a leave-one-out cross-validation strategy was employed in which all normalization, feature selection and classification were performed on all but one training sample to classify the left-out sample. The SVM was trained using a method to generate classification probabilities which then allowed the post-hoc creation of an ROC and the calculation of the AUC measure, as in Study 1. The UAR measure was calculated using the best classification threshold.

3.6. Phonetic analysis

In order to evaluate the prediction that different voice pathologies would have different effects in different phonetic contexts, we took a simple approach and built classifiers for each phone-context separately. There were only 36 phone regions chosen in total, because some phones used by the forced aligner did not occur in all instances of the read passage. In the phonetic evaluation, for each phone type, Summary strategy feature vectors were collated over all segments within the reading passage that were labelled with that phone, and then an SVM classifier was built and validated from the collated data.

3.7. Phonetic fusion

The phone evaluation examined how well regions labelled with the different phones led to pathology classifications. In this regard, each phone context was treated as an independent source of information on the pathology. To fuse the SVM predictions across all phone types, the 36 classification probabilities were fused into a single discriminant using LDA, again with leave-one-out cross validation. In combination with the cross-validation used during training this procedure ensured that both the phone scores and the fusion weights were calculated without reference to the sample under test.

3.8. Baseline results

The main objective of this study was to investigate the benefits of phonetic context in voice disorder discrimination evaluation from a long reading passage, especially as applied to individual diagnoses within the same aetiological type. We compared our proposed system with classification approaches based on vowel productions, as well as continuous speech without phonetic analysis. Baseline results for SD vs VP pathology discrimination are shown in Table 3. For the Functional strategy, UAR performance was around 79% regardless of the speech material used, while performance was a little worse for the Summary strategy. The AUC measure was greatest for the isolated IY vowel, showing slightly better robustness to choice of classification threshold.

3.9. *Phonetic analysis*

To investigate the relative contribution of different phonetic contexts to discrimination of the two pathology classes, we can study the performance of classifiers trained using the Summary strategy for each phone context independently. Table 4 lists some performance figures for SD/VP pathology discrimination for the best and worst single phones. As might have been predicted, all of the best phones are voiced sounds, whilst most of the worst phones are voiceless sounds. Note that the vowels here are found in syllables and are not isolated forms. One possible explanation is that VP imparts a hoarse and breathy quality to voice, which is reflected in more glottal noise during phonation of vowel compared with SD. This difference is visible in Fig. 2. While the best phone models have a greater UAR than the baseline models, the AUC measures for the best phones are similar to those found in the baselines. There is no clear pattern to suggest that better performance comes from phones with a larger number of labelled frames. Exploring why discrimination performance varies across phones for these disorders will be an interesting area of focus in further work.

3.10. *Phonetic fusion*

The classification probabilities were then combined into a single discriminant using LDA. Fig. 3 plots the distribution of the discriminant for the true pathology classes, together with an ROC curve that shows how different choices of threshold for the discriminant leads to changes in Sensitivity and Specificity of the classification. Table 5 provides a confusion matrix for the classification result, using a discriminant threshold of 1.67. The proposed system based on phonetic analysis significantly outperformed the baseline models, obtaining a classification accuracy of 92.1% for SD, 81.8% for VP, and a UAR of 86.96%. This compares to the best baseline UAR of 79.93%. The AUC for the fused system was 0.927 compared to the best baseline of 0.849.

The LDA fusion of phone scores leads to weights for each phone classifier in terms of how much each contribute to a discriminant that separates the SD and VP classes. The phones with the largest and smallest contribution to the discriminant are listed in Table 6, where a positive weight increases the likelihood of a VP classification, and a negative weight increases the likelihood of a SD classification. Interestingly, it is not the phones with best individual classification performance (see Table 4) which have the largest weights. The phones /æ/ and /i/ provide good classification on their own, but contribute little to the discriminant. On the other hand the consonants /z/ and /θ/ provide poor classification on their own but contribute heavily to the discriminant. This suggests that differences between class predictions also contain useful discriminative information, emphasising again that different pathologies have different effects in different phonetic contexts.

Table 3
Baseline results for SD/VP pathology discrimination in terms of AUC and UAR on the recording of a vowel or the whole reading passage. Functional strategy: feature selection of 1000 best features. Summary strategy: median and IQR of LLD features.

Data set	1000 features		252 features	
	AUC	**UAR%**	**AUC**	**UAR%**
IY	0.849	79.39	0.784	73.90
AE	0.818	78.99	0.822	78.20
Passage	0.808	79.93	0.748	75.72

Table 4
Results for best and worst phones in SD/VP pathology discrimination. AUC: area under ROC curve, UAR: unweighted average recall at optimal threshold, Frames: number of 10 ms frames available for that phone label across all recordings.

Best phones				Worst phones			
Phone	**Frames**	**AUC**	**UAR %**	**Phone**	**Frames**	**AUC**	**UAR %**
L /l/	20,259	0.84	86.2	Z /z/	10,985	0.57	63.0
AE /æ/	24,417	0.81	83.3	S /s/	34,623	0.60	64.0
DH /ð/	14,506	0.82	82.3	K /k/	19,304	0.61	66.9
OW /əʊ/	16,397	0.81	82.3	F /f/	17,343	0.66	68.2
IY /i/	17,750	0.84	82.1	T /θ/	40,876	0.66	69.3

3.11. *Discussion*

In this second study, we have presented an automated voice pathology discrimination system based on continuous speech, employing contextualised phonetic analysis of a long reading passage. This system outperforms the baseline models that used the whole recording, whether based on vowels or a read passage, with a 35% reduction in recall error. Moreover, our findings reinforce the hypothesis that voice pathologies influence phonetic contexts in different ways, as phones show different sensitivities for distinct disorder types in the classification. The SD and VP pathologies were selected because of availability, but there are no particular aspects of the method that are specific to these disorders, suggesting that a similar approach might be useful for other pathologies.

In terms of clinical application, post-hoc analysis of the classifier's discriminant shows that the method could lead to a diagnostic test for discriminating between these pathologies that had a sensitivity of about 0.9 for a specificity of about 0.8. These values are similar to many clinical screening tests currently in use [30].

Several limitations regarding the findings are worth noting. First, the relatively small size and the gender imbalance of the pathology samples might have caused problems for classification. A larger, gender-balanced sample would be preferred for future studies. The automated phonetic labelling of the reading passage seemed to work well but relied upon the manual correction of an orthographic transcript to what was actually said. Automation of the generation of the transcript could be a subject for further study, together with an evaluation of the effect of automation on classification accuracy. In addition, phonetic contexts could be considerably expanded, to include, for example, syllable types or prosodic units.

4. Conclusions

In this paper we have presented two complementary studies that tested the hypothesis that voice pathologies can be better discriminated by considering their effects on separate phonetic contexts in continuous speech rather than their average effects on a whole recording. The studies used established machine learning tools and procedures for the extraction of acoustic features (openSMILE), probabilistic classification (SVM), and analysis (LDA). The studies compared contextualised analysis to a baseline and employed careful cross-validation. In study 1 we looked at a large corpus which included a short phrase, while in study 2 we looked at a small corpus that included a long reading passage. The short phrase in study 1 was short enough that it could be annotated using a simple phonetic transcription alignment tool, while the long reading passage in study 2 required the use of a more sophisticated alignment tool that worked from a text transcription. Study 1 looked at two broad pathology classes, while study 2 looked at two narrowly define pathologies. In both studies we have shown improvements in discrimination once phonetic contexts have been taken into account. We suggest that this improvement comes from a sensitivity to the different ways in which voice disorders affect speech production in different articulatory

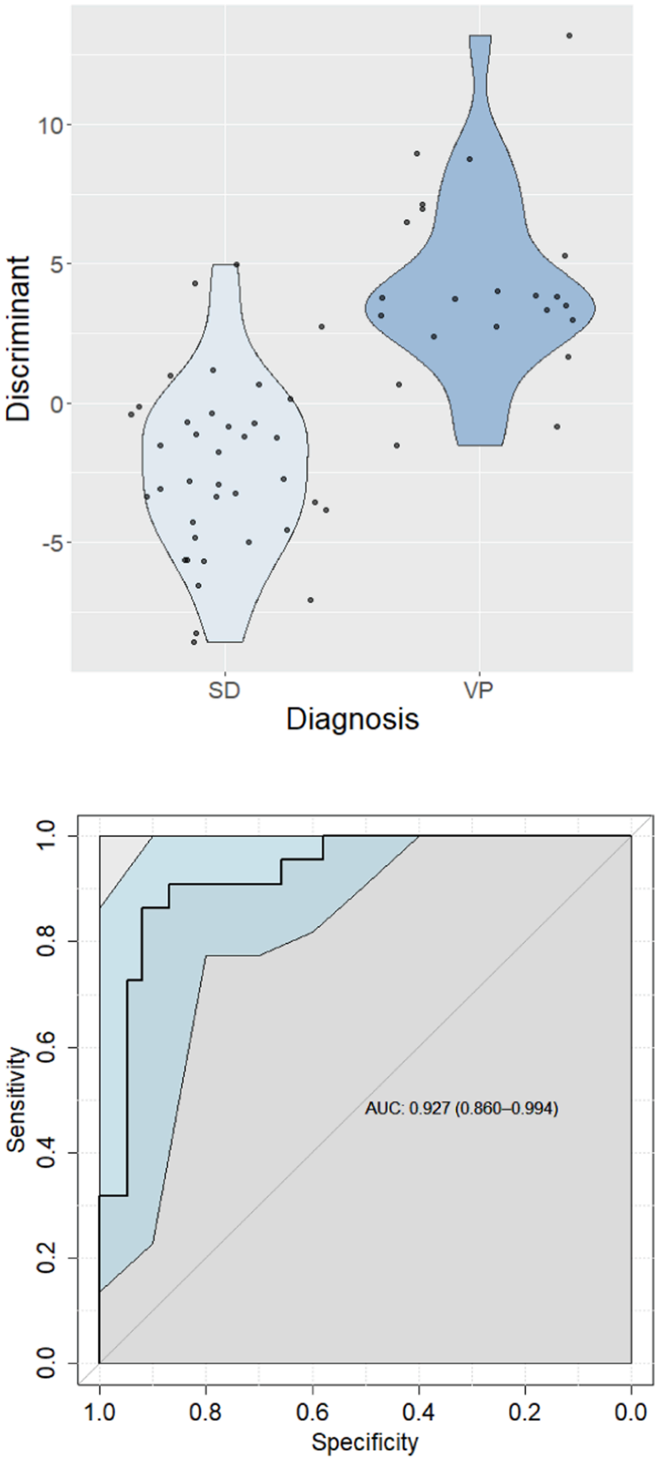

Fig. 3. Distribution of the discriminant from the fused phone classifiers for SD/VP discrimination and an ROC showing the trade-off between sensitivity and specificity.

Table 5
Confusion matrix for voice disorder pathology discrimination using phonetic analysis. UAR = 86.96%.

	SD	VP	Accuracy
SD	35	3	92.1%
VP	4	18	81.8%

Table 6
List of most-weighted and least-weighted phones in voice disorder pathology discriminant (-ve = more SD, +ve = more VP).

Most weighted		Least weighted	
Phone	Weight	Phone	Weight
R /r/	18.7	V /v/	0.1
Z /z/	16.1	AE /æ/	0.3
TH /θ/	−15.4	AW /aʊ/	0.5
AA /ɑ/	13.8	IY /i/	−0.5
L /l/	−10.4	EY /eɪ/	−1.3

contexts as brought out in a continuous speech recording. Importantly, these studies suggest that more sensitive pathology diagnosis might come from designing recording materials to maximise the differences between these contextual effects. While we have not done this yet for our studies, post-hoc analyses of studies like these might reveal which phonetic contexts are most useful for diagnosis and in turn shed light on the interactions between voice disorder, articulation and speech production.

Overall, the performance of our automated method for voice pathology discrimination from audio recordings of continuous speech look promising for a clinical use, achieving sensitivities of 0.9 for levels of specificity which are typical for screening tests. Further work is required to see how robust these methods are to variations in language, accent, recording equipment, reverberation and background noise that will be found in clinical settings. Further research would benefit enormously from a corpus of long continuous speech recordings from a large number of disordered speakers with well-defined pathology labels.

CRediT authorship contribution statement

Mark Huckvale: Conceptualization, Formal Analysis, Methodology. **Zhuoya Liu:** Data Curation, Writing - Revision and Editing. **Catinca Buciuleac:** Data Curation, Writing - Revision and Editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Mark Huckvale reports financial support and article publishing charges were provided by University College London. Mark Huckvale reports a relationship with University College London that includes: employment.

Data availability

The Saarbrucken Voice Database used in study 1 is freely available, the pathological voice recordings used in study 2 are confidential.

Acknowledgements

The authors would like to thank those involved in the creation of the Saarbrücken Voice Database, without which the first study would not have been possible. Thanks to Julian McGlashan for access to the reading passages of the disordered voices. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] J. McGlashan, D. Costello, P.J. Bradley, Hoarseness and voice problems, in: H. Ludman, P.J. Bradley (Eds.), ABC of ear, nose and throat, fifth ed., John Wiley & Sons, 2007.
[2] K. Degila, R. Errattahi, A.E. Hannani, The UCD system for the 2018 FEMH voice data challenge, in: 2018 IEEE International Conference on Big Data, Dec. 2018, pp. 5242-5246.
[3] S.-H. Fang, et al., Detection of pathological voice using cepstrum vectors: A deep learning approach, J. Voice 33 (5) (Sep. 2019) 634–641.

[4] G. Muhammad, et al., Voice pathology detection using interlaced derivative pattern on glottal source excitation, Biomed. Signal Process. Control 31 (Jan. 2018) 156–164.

[5] M. Huckvale, C. Buciuleac, Automated detection of voice disorder in the saarbrücken voice database: Effects of pathology subset and audio materials, in: Proceedings INTERSPEECH 2021–22nd Annual Conference of the International Speech Communication Association, Brno, Czech Republic, Sep. 2021, pp. 1399-1403.

[6] S. Hedge, S. Shetty, S. Rai, T. Dodderi, A survey on machine learning approaches for automatic detection of voice disorders, J. Voice 33 (6) (Nov. 2019) 947. e11–947.e33.

[7] G. Muhammad, M.F. Alhamid, M. Alsulaiman, B. Gupta, Edge computing with cloud for voice disorder assessment and treatment, IEEE Commun. Mag. 56 (4) (Apr. 2018) 60–65.

[8] P. Barche, K. Gurugubelli, A. Kumar Vuppala, Towards automatic assessment of voice disorders: A clinical approach, in: Proceedings INTERSPEECH 2020–21st Annual Conference of the International Speech Communication Association, Shanghai, China, Oct. 2020, pp. 2537-2541.

[9] G. Muhammad, M. Melhem, Pathological voice detection and binary classification using MPEG-7 audio features, Biomed. Signal Process. Control 11 (May. 2014.) 1–9.

[10] Y. Maryn, et al., Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels, J. Voice 24 (5) (Sep. 2010) 540–555.

[11] L. Brinca, et al., The effect of anchors and training on the reliability of voice quality ratings for different types of speech stimuli, J. Voice 29 (6) (Nov. 2015).

[12] H. Cordeiro, C. Meneses, J. Fonseca, Continuous speech classification systems for voice pathologies identification, Doctoral Conference on Computing, Electrical and Industrial Systems 450 (Apr. 2015) 217–224.

[13] S. Wang, C. Wang, C. Lai, Y. Tsao, S. Fang, Continuous speech for improved learning pathological voice disorders, IEEE Open J. Eng. Med. Biol. 3 (Feb 2022) 25–33.

[14] J. Iwarsson, J. Fredsø, Impact of syllable stress and phonetic context on the distribution of intermittent aphonia, Clin. Linguist. Phon. 28 (10) (Oct. 2014) 757–768.

[15] Y.-A. Lien, C.I. Gattuccio, C.E. Stepp, Effects of phonetic context on relative fundamental frequency, J. Speech Lang. Hear. Res. 57 (4) (Aug. 2014) 1259–1267.

[16] J. Kane, M. Aylett, I. Yanushevskaya, C. Gobl, Phonetic feature extraction for context-sensitive glottal source processing, Speech Comm. 59 (Apr. 2014) 10–21.

[17] J. Laver, The phonetic description of voice quality, Cambridge University Press, Cambridge, 1980.

[18] L. Moro-Velazquez, et al., Phonetic relevance and phonemic grouping of speech in the automatic detection of Parkinson's Disease, Sci. Rep. 9 (1) (Dec. 2019) 19066.

[19] M.G. Tulics, K. Vicsi, Phonetic-class based correlation analysis for severity of dysphonia, in: 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Sep. 2017, pp. 21-26.

[20] M. Putzer, W. Barry, "Saarbrucken Voice Database", Institute of Phonetics, Univ. of Saarland, Accessed March 2021 from http://www.stimmdatenbank.coli.uni-saarland.de/.

[21] M. Putzer, J. Koreman, A German database of pathological vocal fold vibration, Phonus 3 Institute of Phonetics, University of the Saarland, 1997, pp. 143-153.

[22] UCL Division of Psychology and Language Sciences, "Speech Filing System", Version 4.10, Retrieved February 2023 from https://www.phon.ucl.ac.uk/resource/sfs/.

[23] F. Eyben, M. Wollmer, B. Schuller, Opensmile: the Munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1459–1462.

[24] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi et al., The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism, in: Proceedings INTERSPEECH 2013-14th Annual Conference of the International Speech Communication Association, Lyon, France, Aug. 2013, pp.148-152.

[25] R. Peterson, "bestNormalize: Normalizing transformation functions", Version 1.7.0, Retrieved March 2021 from https://cran.r-project.org/web/packages/bestNormalize/.

[26] D. Meyer et al., "E1071: Miscellaneous functions of department of statistics TU Wien", Version 1.7-6, Retrieved March 2021 from https://cran.r-project.org/web/packages/e1071/.

[27] E. Smith, et al., Spasmodic dysphonia and vocal fold paralysis: Outcomes of voice problems on work-related functioning, J. Voice 12 (2) (1998) 223–232.

[28] M. McAuliffe et al., Montreal forced aligner: trainable text- speech alignment using Kaldi, in: Proceedings INTERSPEECH 2017–18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, Aug. 2017, pp. 498-502.

[29] B. Schuller et al., The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism, in: Proceedings INTERSPEECH 2013-14th Annual Conference of the International Speech Communication Association, Lyon, France, Aug. 2013, pp. 148-152.

[30] M. Leeflang, Systematic reviews and meta-analyses of diagnostic test accuracy, Clin. Microbiol. Infect. 20 (2) (2014) pp.

[31] Z. Liu, M. Huckvale, J., McGlashan, Automated Voice Pathology Discrimination from Continuous Speech Benefits from Analysis by Phonetic Context, in Proceedings INTERSPEECH 2022-23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, Sep. 2022, pp. 2158-2162.