Interspeech 2022
18-22 September 2022, Incheon, Korea

Automated Voice Pathology Discrimination from Continuous Speech Benefits from Analysis by Phonetic Context

Zhuoya Liu [1], *Mark Huckvale*[1], *Julian McGlashan*[2]

[1] Speech, Hearing and Phonetic Sciences, University College London, U.K.
[2] ENT Department, Nottingham University Hospitals, U.K.

zhuoya.liu.20@alumni.ucl.ac.uk, m.huckvale@ucl.ac.uk,
julian.mcglashan@nottingham.ac.uk

Abstract

In contrast to previous studies that look only at discriminating pathological voice from the normal voice, in this study we focus on the discrimination between cases of spasmodic dysphonia (SD) and vocal fold palsy (VP) using automated analysis of speech recordings. The hypothesis is that discrimination will be enhanced by studying continuous speech, since the different pathologies are likely to have different effects in different phonetic contexts. We collected audio recordings of isolated vowels and of a read passage from 60 patients diagnosed with SD (N=38) or VP (N=22). Baseline classifiers on features extracted from the recordings taken as a whole gave a cross-validated unweighted average recall of up to 75% for discriminating the two pathologies. We used an automated method to divide the read passage into phone-labelled regions and built classifiers for each phone. Results show that the discriminability of the pathologies varied with phonetic context as predicted. Since different phone contexts provide different information about the pathologies, classification is improved by fusing phone predictions, to achieve a classification accuracy of 83%. The work has implications for the differential diagnosis of voice pathologies and contributes to a better understanding of their impact on speech.

Index Terms: voice pathology discrimination, phonetic context, continuous speech, machine learning

1. Introduction

The human voice production system can become impaired in multiple ways involving structural, neurogenic, inflammatory or muscle tension imbalance [1]. Differentiation between types of disorders by subjective auditory assessments of clinicians is difficult because of similarities in auditory effect, and diagnostic reliability is highly influenced by clinician training, background, and experience [2]. Instrumental methods are based on endoscopic examination of the larynx and for voice assessment involving Acoustic and Electroglottographic (EGG) analysis are available but only in specialized centers. 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 [3, 4, 5]. Although there are many such studies focusing on contrasting pathological voice from neurotypical voice (see [6] for a survey), few studies in the past few decades have investigated differential diagnosis of voice pathologies [7, 8, 9]. However, when voice problems are concerned, it is essential to determine the underlying causes to provide appropriate and effective medical treatment.

The most common approach to automated voice disorder assessment has been to use sustained vowel productions (e.g., [ɑ], [e], and [i]) instead of continuous or connected speech. Although sustained vowels have long been used by clinicians to assess voice, they lack ecological validity [10]. Since continuous speech requires the exercise of more laryngeal functions it seems likely that this style would better expose voice disorders. Recent studies of automated assessment have indeed shown that pathology detection can be better from recordings of continuous speech than from vowels [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 parts of a passage will be more informative than others about the voice pathology, and so the computing of averages across a whole recording will dilute features that might be very useful for discrimination of pathologies.

The use of a fixed reading passage for vocal assessment enables researchers to analyze voice characteristics according to known phonetic and phonological contexts. This analysis of voice by phonetic context could provide key discriminating information that is not present in the averages computed over whole recordings, since the production of different phones is associated with different vocal tract and laryngeal configurations and activities. For example, the starting and stopping of voicing in plosives, or the aerodynamic interactions between vocal tract and voicing in voiced fricatives might expose disorders in ways not obvious in a sustained vowel. Previous studies have shown how in dysphonic voice with vocal fold thickening, unstressed syllables are more likely to be produced with insufficient subglottal pressure realizing aphonia [14]. The variation of voice quality with phonetic context in the normal voice 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] or 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). However, this idea has not yet been applied to voice pathology discrimination.

In this paper, we propose a new phonetic analysis method for automatic voice pathology discrimination from a clinical perspective. To the best of our knowledge, this is the first

10.21437/Interspeech.2022-10154

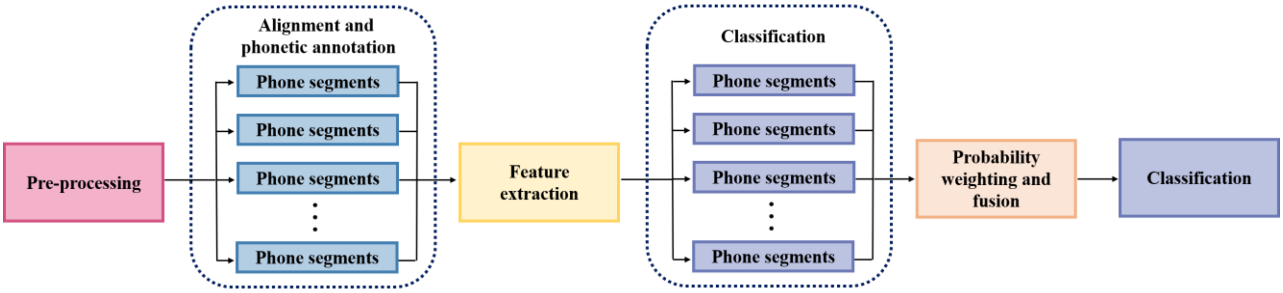

Figure 1: *The framework of proposed method for automated voice pathology discrimination based on phonetic context.*

attempt to automatically discriminate voice pathologies on the basis of continuous speech. Our hypothesis is that different voice pathologies will cause measurably different acoustic variations in different phonetic contexts, which can be used for discriminating them. The work has the following novel elements: (i) application of automatic forced alignment and phonetic annotation techniques, which enables the segmentation of a recording of a passage according to phonetic contexts; (ii) investigation of how different phonetic contexts affect the classification of two different voice pathologies, serving to determine more appropriate speech tasks for classifying pathologies; and (iii) introduction of a novel phonetic fusion method that can significantly improve voice pathology classification accuracy.

2. Materials and Methods

Our proposed method aims to analyze phonetic context to improve the automatic pathology classification from continuous speech, using a widely available feature toolkit (OpenSMILE) [20] and classifier (Support Vector Machine, SVM). The framework of the method is illustrated in Figure 1.

2.1. 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). 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 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 [21].

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 samp/sec 16-bit, while some were at 22,050 samp/sec. Only the recorded speech signals were used in this study, and the EGG recordings will be analyzed in a later study.

2.2. Audio pre-processing

For the baseline trials, two types of vowel-sound extracts were segmented from recordings of the production of sustained vowels collected in another assessment and the whole recording of passage reading was used:

- IY: instance of an [i] vowel spoken on a low pitch.
- AE: instance of an [æ] vowel in the isolated word "sat".
- Passage: reading of "Arthur the Rat" passage. The average duration of SD recordings was 149 seconds, and the average duration of VP recordings was 141 seconds.

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 32,000 samples/sec.

2.3. 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 editions 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 several words. The orthographic transcript was then aligned with the speech audio using the Montreal forced aligner [22]. 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. Figure 2 shows the examples of the automatic alignment and annotation of the word 'Arthur'.

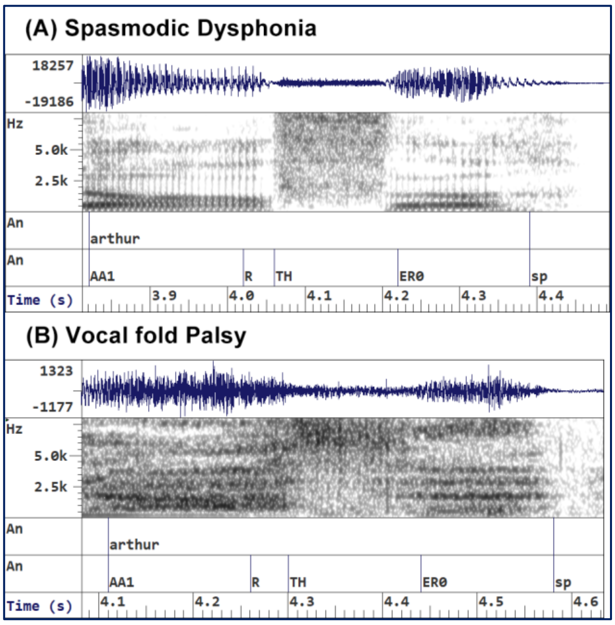

Figure 2: *Examples of automatic alignment and annotation. Voice recordings are from a speaker with SD (above) and a speaker with VP (below) producing the word 'Arthur'.*

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

2.4. Feature extraction

The OpenSMILE analysis system was used to extract features for processing. We adopted two strategies for summarizing features across recordings. In the Functional strategy, we used the large set of summary functionals found in the COMPARE13 configuration of OpenSMILE [23], which delivered 6373 features per recording. In the Summary strategy, we used the COMPARE13 low-level descriptors (LLD) configuration delivering 126 features per 10ms frame and then computed the median and inter-quartile range of each LLD to give 252 features for each specified region. This latter approach allowed us to label each frame with the active phone so that summaries could be produced for the whole recording or specific phonetic regions.

2.5. Classification

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 normalization of features was performed 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. This present study used a Support Vector Machine (SVM) classifier from the e1071 package for R [24]. A radial basis function kernel was selected with a cost parameter C=2.

2.6. 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 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.

2.7. Phonetic fusion

The phone evaluation examined how well regions labelled with the different phones led to pathology classifications. In this regard, each phone region was treated as an independent source of information on the pathology. This suggests that improved classification performance can be obtained by fusing the classification predictions made from different phones. To fuse the predictions, the SVM classifiers were again trained for each labelled phone region, but in such a way as to provide a pseudo "probability" of classification. This then generated a vector of 36 scores for each recording—representing the probability that the recording came from an SD case assessed by each phone type. Then, score fusion was performed by computing the weighting of the probabilities that best discriminated the two pathologies. This was implemented using linear discriminant analysis (LDA), again with leave-one-out cross-validation. This cross-validation procedure ensured that both the phone scores and the fusion weights were calculated without reference to the sample under test.

3. Results

3.1. Baseline results

The main objective of this paper is to investigate the benefits of phonetic context in voice disorder discrimination evaluation. We compared our proposed system with classification approaches based on vowel productions as well as continuous speech without phonetic analysis. Baseline results for pathology discrimination are shown in Table 1. Values are unweighted average recall (UAR). It is noteworthy that the UARs for vowel sounds were slightly higher compared with passage reading. This difference might be due to the problem described in the introduction of this paper—some small but meaningful changes can be lost due to the overall variability when the features are averaged across the recording.

Table 1: *Baseline results for voice disorder pathology discrimination in terms of UAR on the whole recording. Functional strategy: feature selection of 1000 best features. Summary strategy: median and IQR of LLD features.*

Data set	Functional Strategy (1000 features)	Summary Strategy (252 features)
IY [i]	75.44%	73.36%
AE [æ]	73.49%	70.53%
Passage	71.65%	70.33%

3.2. Phonetic analysis

The results of the phonetic evaluation show that the performance for different phones differed significantly, confirming that phonetic context is important for voice disorder classification. Table 2 includes pathology discrimination accuracy for the best and worst phones. Most of the best phones are vowel sounds, whilst all worst phones are voiceless consonant sounds. Note that the vowels here are found in syllables and are not isolated forms. There is no clear pattern to suggest that better performance comes from a larger number of labelled frames. Further analysis is required to investigate the reasons why particular phone types aid discrimination, and to also assess their statistical significance.

Table 2: *Results for best and worst phones in voice disorder pathology discrimination. Frame represents the number of 10ms frames used for the classification across all recordings*.

Best phones			Worst phones		
Phone	Frame	UAR %	Phone	Count	UAR %
AH [ʌ]	48227	78.5	P [p]	3275	59.6
EH [ɛ]	19288	78.5	F [f]	17343	58.6
EY [eɪ]	24475	77.5	K [k]	19304	58.6
DH [ð]	14506	77.2	CH [tʃ]	5590	47.5
ER [ɜ]	17317	75.8	TH [θ]	7737	45.6

3.3. Phonetic fusion

The LDA fusion of phone scores leads to weights for each phone type in terms of how much they contribute to a discriminant that separates the SD and VP classes. The weightings given to the most "SD sensitive" and most "VP sensitive" phone classifications are summarized in Table 3. The results suggest that SD and VP affect phonetic contexts in different ways. We noticed that nasal sounds (i.e. NG [ŋ], M [m], and N [n]) contributed particularly to the discrimination, while vowel sounds such as UH [ʊ], OY [ɔɪ], AO [ɔ], AA [ɑ], and EH [ɛ] only had small weights, with absolute values below 1.

Table 3: *Results for most "SD sensitive" and most "VP sensitive" phones in voice disorder pathology discrimination (-ve = more SD, +ve = more VP).*

Most SD sensitive		Most VP sensitive	
Phone	Weight	Phone	Weight
IH [ɪ]	-11.662	NG [ŋ]	13.576
Z [z]	-11.617	K [k]	11.714
B [b]	-9.252	N [n]	10.384
M [m]	-7.226	AY [aɪ]	8.596
EY [eɪ]	-5.948	CH [tʃ]	8.179

Figure 3 plots the distribution of the discriminant for the true pathology classes, and Table 4 provides a confusion matrix for the classification result, using a discriminant threshold of zero. The proposed system based on phonetic analysis significantly outperformed the baseline models, obtaining a classification accuracy of 89.5% for SD, 77.3% for VP, and a UAR of 83.37%. This compares to the best baseline UAR of 75.44%.

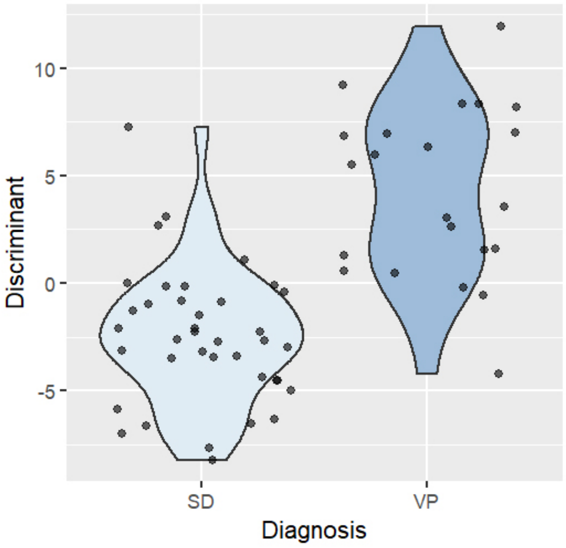

Figure 2: *Distribution of the discriminant for the true SD and VP pathology classes.*

Table 4: *Confusion matrix for voice disorder pathology discrimination using phonetic analysis. UAR=83.37%.*

	SD	VP	Accuracy
SD	34	4	89.5%
VP	5	17	77.3%

4. Discussion and Future Work

In this paper, we presented an automated voice pathology discrimination system based on continuous speech, employing a novel phonetic context analysis method. This system outperforms the baseline models that used the whole recording, whether based on vowels or a read passage, with a 32% 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 is specific to these disorders, suggesting that a similar approach might be useful for other pathologies We believe that the present work not only provides important implications for the future design of effective discrimination systems as well as vocal tasks but also contributes to a better understanding of the mechanisms of voice pathologies.

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. Finally, future work should also investigate the benefits that arise from joint analysis of the speech audio with the EGG signals.

5. References

[1] J. McGlashan, D Costello, and P.J. Bradley, "Hoarseness and voice problems" in H. Ludman and P. J. Bradley (Eds.), *ABC of ear, nose and throat 5th Edition*. John Wiley & Sons, 2007.

[2] K. Degila, R. Errattahi, and A.E. Hannani, "The UCD system for the 2018 FEMH voice data challenge," *2018 IEEE International Conference on Big Data*, pp. 5242-5246, Dec. 2018.

[3] S.-H. Fang, *et al.*, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *Journal of Voice*, vol. 33, no. 5, pp. 634–641, Sep. 2019.

[4] G. Muhammad *et al.*, "Voice pathology detection using interlaced derivative pattern on glottal source excitation," *Biomedical Signal Processing and Control*, vol. 31, pp. 156-164, Jan. 2018.

[5] M. Huckvale and 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, and T. Dodderi, "A survey on machine learning approaches for automatic detection of voice disorders," *Journal of Voice*, vol.33, no. 6, pp. 947.e11-947.e33, Nov. 2019.

[7] G. Muhammad, M. F. Alhamid, M. Alsulaiman, and B. Gupta, "Edge computing with cloud for voice disorder assessment and treatment," *IEEE Communications Magazine*, vol. 56, no. 4, Apr. 2018, pp. 60-65.

[8] P. Barche, K. Gurugubelli, and 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 and M. Melhem, "Pathological voice detection and binary classification using MPEG-7 audio features," *Biomedical Signal Processing and Control*, vol. 11, pp. 1-9, May. 2014.

[10] Y. Maryn *et al.*, "Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels," *Journal of Voice*, vol. 24, no. 5, pp. 540-555, Sep. 2010.

[11] L. Brinca *et al.*, "The effect of anchors and training on the reliability of voice quality ratings for different types of speech stimuli," *Journal of Voice*, vol. 29, no.6, pp. 776.e777-776.e714, Nov. 2015.

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

[13] S. Wang, C. Wang, C. Lai, Y. Tsao, and S. Fang, "Continuous speech for improved learning pathological voice disorders," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 3, Feb, 2022, pp. 25-33.

[14] J. Iwarsson and J. Fredsø, "Impact of syllable stress and phonetic context on the distribution of intermittent aphonia," *Clinical Linguistics & Phonetics*, vol. 28, no.10, pp. 757-768, Oct. 2014.

[15] Y.-A. Lien, C. I. Gattuccio, and C. E. Stepp, "Effects of phonetic context on relative fundamental frequency," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 4, pp. 1259-1267, Aug. 2014.

[16] J. Kane, M. Aylett, I. Yanushevskaya, and C. Gobl, "Phonetic feature extraction for context-sensitive glottal source processing," *Speech Communication*, vol. 59, pp. 10-21, Apr. 2014.

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

[18] L. Moro-Velazquez *et al.*, "Phonetic relevance and phonemic grouping of speech in the automatic detection of Parkinson's Disease," *Scientific Reports*, vol. 9, no.1, pp. 19066, Dec. 2019.

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

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

[21] E. Smith *et al.,* "Spasmodic dysphonia and vocal fold paralysis: Outcomes of voice problems on work-related functioning," *Journal of Voice*, vol.12, no.2, pp.223-232, 1998.

[22] 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.

[23] 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.

[24] D. Meyer *et al.*, "e1071: Misc Functions of the Department of Statistics," https://cran.r-project.org/web/packages/e1071/