INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Recognising Emotions in Dysarthric Speech Using Typical Speech Data

Lubna Alhinti[1], *Stuart Cunningham*[2,3], *Heidi Christensen*[1,3]

[1]Department of Computer Science, University of Sheffield, United Kingdom
[2]Health Sciences School, University of Sheffield, United Kingdom
[3] Centre for Assistive Technology and Connected Healthcare (CATCH), Sheffield, United Kingdom
{laalhinti1,s.cunningham,heidi.christensen}@sheffield.ac.uk

Abstract

Effective communication relies on the comprehension of both verbal and nonverbal information. People with dysarthria may lose their ability to produce intelligible and audible speech sounds which in time may affect their way of conveying emotions, that are mostly expressed using nonverbal signals. Recent research shows some promise on automatically recognising the verbal part of dysarthric speech. However, this is the first study that investigates the ability to automatically recognise the nonverbal part. A parallel database of dysarthric and typical emotional speech is collected, and approaches to discriminating between emotions using models trained on either dysarthric (speaker dependent, *matched*) or typical (speaker independent, *unmatched*) speech are investigated for four speakers with dysarthria caused by cerebral palsy and Parkinson's disease. Promising results are achieved in both scenarios using SVM classifiers, opening new doors to improved, more expressive voice input communication aids.

Index Terms: Dysarthria, emotion recognition, communication aids

1. Introduction

Dysarthria is a neurological disorder that interferes with articulation, respiration, phonation, and resonance. There are a number of underlying causes of dysarthria including traumatic brain injury, stroke, Parkinson's disease (PD), cerebral palsy, and multiple sclerosis [1]; it is one of the most commonly acquired speech disorders [2]. The speech intelligibility of people with dysarthria is affected and conveying emotions in their speech, in a way that can be understood clearly by listeners, might be difficult due to their prosodic and phonological dimension limitations. As a result, people with dysarthria may experience a number of complications including serious communication problems, that can impact their lives negatively and lead to social isolation and depression.

Recent research efforts have concentrated on finding ways to automatically recognise the *verbal* part of dysarthric speech. Developing dysarthric automatic speech recognition (ASR) is considered to be a challenging task due to the intra- and inter-speaker variability in dysarthric speech and the difficulty of obtaining suitable data, that is, matched and in sufficient quantity [3, 4]. Researchers have been working on developing and improving dysarthric ASR systems and employing different techniques to overcome the challenges such as using adaptation techniques and speaker dependent models [5, 6, 7, 3]. None of those studies have focused on recognising any of the non-verbal, precisely paralinguistic, information such as the emotional state of the speaker.

The ability to recognise emotion is an important component in social interaction. People can express their emotions through a number of different modalities including speech, voice characteristics, facial expressions, and gestures. Automatic speech emotion recognition (SER) has gained a lot of interest due to its increasingly important role in many fields including human-robot interaction, call centers, gaming, and education. In addition to its applications in healthcare and assistive technology. There is a huge body of work done by researchers to automatically recognise emotions from human speech including recognising discrete emotions, recognising positive and negative emotions, and recognising nonverbal sounds such as cries and laughter [8, 9].

Although emotions, to the best of our knowledge, have never been investigated in dysarthric speech, a number of studies investigated the prosodic and phonatory features of dysarthric vocalisation. The ability of people with severe spastic dysarthria, caused by cerebral palsy, to signal question-statement contrast was investigated [10]. 87% classification accuracy of questions versus statements was achieved in the subjective evaluation using only prosodic cues. The strategies of speakers with dysarthria of signaling the contrast compared to typical speakers were also investigated [11]. It was found that while typical speakers primarily rely on fundamental frequency (F0) and duration cues; speakers with dysarthria additionally use intensity. The ability of people with severe dysarthria caused by cerebral palsy to convey information using pitch and duration cues was also investigated [12]. All 8 speakers were able to control duration while their ability to control pitch varies. Likewise, using prosodic characteristics, the signalling of question-statement contrast in Mandarin, German, and Cantonese languages of people with hypokinetic dysarthria caused by PD was investigated and compared to healthy control speakers [13, 14, 15]. Although speakers with dysarthria can differ from control speakers in their prosodic characteristics, they were able to signal question-statement contrast like healthy control speakers except for the Mandarin speakers. High variability among speakers was observed in some of these studies in terms of their ability to acoustically signal questions. The prosodic and acoustic characteristics such as F0, speech rate, and intensity of people with dysarthria caused by PD were also investigated and compared to healthy control speakers [16, 17, 18, 19, 20, 21].

There is a lack of consistency in some of the reported results as some studies show differences in some of these features between the two groups while other reported no differences. This indicates the high variability in the acoustic characteristics of people with dysarthria. Nevertheless, it shows they may have enough control to convey emotions and show intentions in their speech which opens up new horizons for improving communication aids. An example of such a communication aid is the voice-input-voice-output communication aid (VIVOCA) which is a communication aid that recognises a person's (dis-

Copyright © 2020 ISCA

http://dx.doi.org/10.21437/Interspeech.2020-1825

ordered) speech and reproduces it in a synthesised, more intelligible, voice is a form of a voice-driven communication aid [22]. Adding expressiveness to the synthesised voice would enhance the users' communication experience and social relationship, especially since many communication aid users prefer to use their residual voice when they communicate. This paper investigates the first step towards this goal which is the ability to automatically recognise emotions in the speech of people with dysarthria by using a novel collected database. Employing state of the art SER methods directly to the domain of dysarthric emotion recognition is challenging due to several reasons. Mainly, these methodologies require a large amounts of data to perform well. However, collecting large amounts of dysarthric emotional speech data is more challenging than emotional typical speech. Also, the high inter-speaker variability found in this group of speakers in comparison to typical speakers poses another challenge [4]. Since the use of typical speech data to boost the performance on dysarthric speech data has been applied successfully to the domain of dysarthric ASR [23], we investigate whether the dysarthric emotion recognition model would benefit from being trained on typical speech data and to what extent you can accurately classify emotions in dysarthric speech using typical speech models. In other words, investigating whether people with dysarthria share some similarities with typical speakers while expressing emotions.

2. Data

Experiments reported in this paper are based on a newly collected database of parallel recordings of dysarthric and typical speech. Over the past decades, there has been considerable debate over how best to collect the emotional speech for such collections. Either using natural, elicited (induced), or acted speech. Adopting the natural methodology will not be appropriate in establishing this database as determining the underlying emotion would be even more challenging than for typical speech, which in turn will add more ambiguity to this unexplored domain. The acted methodology will also not be appropriate as there are no, or very hard to find, actors with dysarthric speech. The elicitation approach was therefore chosen in which very short video clips of emotion stimuli is used. This follows standard protocols for recording such databases. The video clips were adopted from those used by [24] when recording the SAVEE database, a British English typical speech audio-visual emotional database. The video clips originate from popular movies and series.

All participants were British English speakers over 18 and with no known cognitive problems and no known literacy difficulties. Participants were not professional actors. Table 1 outlines the speakers' details. Recordings of speakers with PD were taken while they were under the anti-Parkinsonian medications effect.

The database recordings took place in a professional recording studio at the University of Sheffield. Recordings were obtained from people with typical and dysarthric speech (caused by either cerebral palsy or PD). Figure 1 shows the data capture physical setup. Both audio and video were recorded; here we only use the audio recordings. Speakers were recorded individually using a Marantz PMD 670 recorder in mono and sampled at 16,000 Hz. The microphone was placed 50 cm from the speaker. The prompting material was displayed on a 13 inch Macbook Air and placed on a table 1 meter from the speaker.

Speakers were recorded expressing all the six basic emotions: angry, happy, sad, surprise, fear, and disgust in addition to neutral. Given that this is a first of its kind study, starting with a smaller set can provide the base for a more focused initial exploration of the problem. Also, informed by a survey the authors have conducted in order to understanding barriers to communicating emotion by people with dysarthria, anger, happiness and sadness were chosen in addition to neutral as a baseline condition [25].

Table 1: *Speaker details*

Speakers with dysarthria				
Type of dysarthria	**Speaker**	**Gender**	**Age**	**Dysarthria severity */ time diagnosed**
Spastic dysarthria (cerebral palsy)	DS01F	Female	65 years	Severe / from birth
Hypokinetic dysarthria (PD)	DS02F	Female	71 years	Mild / 10 years
	DS04F	Female	68 years	Mild-to-moderate / 10 years
	DS03M	Male	66 years	Moderate / 9 years
Speakers with typical speech				
Gender	**Number of Speakers**	**Age**		
		Mean	**SD**	**Range**
Female	9	34.00	13.26	20-56
Male	12	35.67	16.81	19-70

* Dysarthria severity levels in the table are informal judgments by the authors.

The text material (also adopted from [24]) contains TIMIT sentences of common, emotion-specific, as well as generic sentences for each emotion. Long sentences were excluded and the final set includes 10 sentences per emotion and 20 neutral sentences. This gives a total of 50 sentences per speaker for the selected emotions. The stimuli presentation consisted of three main stages: task presentation (states the emotion), emotive video presentation, and sentence presentation. Each emotion's sentences were split into two equal sets. Each round begins with a neutral set followed by one set from each emotion, giving a total of two rounds to record all the sentences. This division procedure was applied to help in avoiding bias caused by speakers' fatigue. All speakers finished their recordings in one session. The data will be made publicly available in the near future with more detailed specifications.

For this experiment, the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) was chosen. It contains spectral, prosodic, cepstral, and voice quality information such as F0, jitter, shimmer, harmonic differences, and MFCC for a total of 88 features [26]. The eGeMAPS has been widely used as a benchmark for emotion classification studies [27, 28, 29]. Features were automatically extracted using the openSMILE toolkit [30] using the default parameters [26]. All features were standardised so that they have zero mean and unit variance

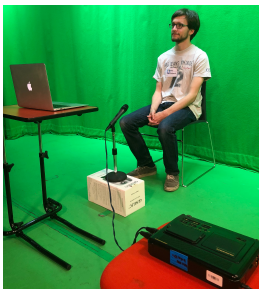


Figure 1: *Data capture physical setup.*

3. Classification results

The aim of this experiment is to investigate approaches to automatically recognise emotions from dysarthric speech including to what degree this is achievable using a model trained on typical speech only. Therefore, two classification approaches were performed: 'speaker dependent' (SD) where the model is trained and tested using the target speaker's voice characteristics and 'speaker independent' (SI) where the model is trained using typical speech from a number of speakers and tested on a dysarthric speaker's voice. An SVM classifier with radial basis function (RBF) kernel was used for all classification tasks. The regularisation parameter (C) and the gamma coefficient of the kernel were set to 5 and 0.01, respectively. The Scikit-learn package was used to train the classifier [31]. SVMs are widely used for tasks involving pathological speech as they are particularly well-suited to sparse data domains.

3.1. Training and test data

Due to the relatively small dataset in the SD classification tasks, a 5-fold cross validation technique was applied to increase the reliability of the results. The splits in each fold preserved the samples' distribution in each emotion. Since it is widely believed that females express emotions differently than males do, we took this into consideration while choosing the training and testing datasets for the SI classification tasks [32]. Thus, all female typical speakers were used as training data when the target speaker was a female speaker with dysarthria. The same goes for male speakers. The recognition performance is reported for each speaker separately taking into consideration the inter-speaker variability in dysarthric speech. This also helps in establishing clear baseline results for each speaker.

3.2. Recognising four emotional states

The aim of the first experiment is to recognise four emotional states and compare the performance of the two classification approaches: SD and SI. Table 2 presents the recognition results for all speakers. The accuracy results of both experiments are illustrated in Figure 2. Since the two classification tasks have a big difference in the amount of training data, we repeated the SI classification task using only half of the training data but found that this does not have a significant impact on the classification results.

From Table 2, it is observed that for all speakers 'anger' is never confused with 'sad' and 'sad' is rarely confused with 'anger'. For speaker DS01F, 'anger' is mostly confused with 'happy' and this confusion increases when the model is trained on typical speech. It is also observed that 'neutral' is mostly confused with 'happy' for speaker DS01F when the model is trained on typical speech. A big improvement is achieved for classifying 'happy' for speaker DS04F after training the model on typical speech. For speaker DS03M, 'happy' is most likely to be confused with 'anger' when the model is trained on typical speech. For all speakers, except speaker DS01F, 'anger' and 'neutral' are mostly considered as non confusing pairs. It is also observed that 'neutral' is mostly confused with 'sad'. Similar results were observed on typical speech in [33, 34]. In general, the results of both classification approaches are above chance performance for all speakers, with better performance obtained from the targeted SD models.

The relatively good results when testing the dysarthric speakers on the SI models are encouraging. This indicates that a good level of typical-like emotion specific information is being successfully expressed. In comparison, testing the typical speakers on the SI models using leave-one-speaker-out approach, where the model is trained on all typical speakers data except one speaker who was held as a test set, achieved an average accuracy of 59.81%, showing that this is a difficult task even when staying within the typical domain.

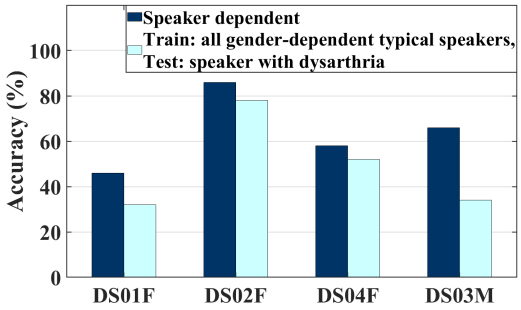


Figure 2: *Recognition results with different training data*

Table 2: *The results of recognising 4 emotions. (rows = real values and columns = predicted values).*

		Train	Speaker dependent				All gender-dependent typical speaker			
			Ang	Hap	Sad	Neu	Ang	Hap	Sad	Neu
Test	DS01F	Ang	4	3	0	3	1	6	0	3
		Hap	3	2	3	2	0	6	0	4
		Sad	1	0	5	4	0	4	0	6
		Neu	3	3	2	12	5	6	0	9
	DS02F	Ang	9	1	0	0	10	0	0	0
		Hap	0	8	0	2	0	8	1	1
		Sad	0	1	8	1	0	0	10	0
		Neu	0	0	2	18	0	1	8	11
	DS04F	Ang	8	2	0	0	8	2	0	0
		Hap	5	1	2	2	1	7	0	2
		Sad	0	1	5	4	1	2	7	0
		Neu	0	2	3	15	1	4	6	9
	DS03M	Ang	8	1	0	1	9	1	0	0
		Hap	2	5	1	2	8	2	0	0
		Sad	1	1	4	4	6	3	1	0
		Neu	1	0	3	16	2	7	6	5

3.3. Recognising two emotional states

The aim of this experiment is to distinguish pairs of emotions using the SI classification approach. The results of recognising the following pairs: anger/happy, anger/sad, and happy/sad for all speakers with dysarthria are presented in Table 3 and Figure 3.a. High classification accuracies are obtained for the two female speakers DS02F and DS04F are very good. This is not the case for speakers DS01F and DS03M where most of the results are at chance level or a little bit above chance except when recognising anger/sad for speaker DS01F, where higher accuracy results are achieved. The most likely explanation is that speakers DS01F and DS03M are more severely dysarthric speakers than the other two speakers.

3.4. Discriminating emotional speech from neutral speech

A third experiment was conducted using the same setup used in the above experiment to see how well dysarthric emotional speech can be distinguished from dysarthric neutral speech. Table 4 and Figure 3.b present the results of distinguishing the fol-

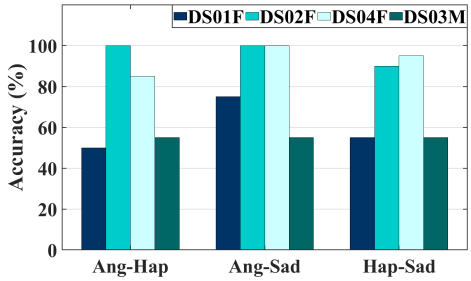
Table 3: *The results of recognising 2 emotional states when trained using gender-based emotional typical speech. (rows = real values and columns = predicted values).*

	DS01F		DS02F		DS04F		DS03M	
	Ang	Hap	Ang	Hap	Ang	Hap	Ang	Hap
Ang	0	10	10	0	8	2	9	1
Hap	0	10	0	10	1	9	8	2
	Ang	Sad	Ang	Sad	Ang	Sad	Ang	Sad
Ang	7	3	10	0	10	0	10	0
Sad	2	8	0	10	0	10	9	1
	Hap	Sad	Hap	Sad	Hap	Sad	Hap	Sad
Hap	10	0	8	2	9	1	10	0
Sad	9	1	0	10	0	10	9	1

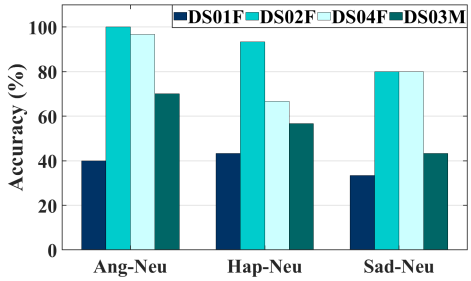
Table 4: *Results of discriminating emotional speech from neutral speech when trained using gender-based emotional typical speech. (rows = real values and columns = predicted values).*

	DS01F		DS02F		DS04F		DS03M	
	Ang	Neu	Ang	Neu	Ang	Neu	Ang	Neu
Ang	4	6	10	0	10	0	10	0
Neu	2	8	0	20	1	19	9	11
	Hap	Neu	Hap	Neu	Hap	Neu	Hap	Neu
Hap	7	3	9	1	8	2	10	0
Neu	14	6	1	19	8	12	13	7
	Sad	Neu	Sad	Neu	Sad	Neu	Sad	Neu
Sad	10	0	10	0	5	5	8	2
Neu	20	0	6	14	1	19	15	5

lowing pairs: anger/neutral, happy/neutral, and sad/neutral. The classifications accuracy of speaker DS01F are below chance level performance for all three pairs of emotions. For the other speakers, the results of discriminating emotional from non emotional speech are mostly very good. On average, it is found that the easiest pair to discriminate is anger/neutral. Similar results were reported on typical speech in [33, 34].



(a) Results of recognising two emotional states



(b) Results of discriminating emotional speech from neutral speech

Figure 3: *The result of recognising pairs of emotions*

4. Discussion and conclusion

Given the nature of the dysarthric speech and its phonological and prosody dimensions limitations, the experiments in this work were conducted to investigate i) the feasibility of automatically recognising emotions from dysarthric speech, ii) whether there are similarities between emotional typical speech and emotional dysarthric speech, and iii) what emotions in the dysarthric speech are found to be close to each other (confusing) in the chosen feature space. We demonstrated that dysarthric speech emotion recognition could be possible. In fact, the results of recognising four emotional states were above chance performance for all speakers including, DS01F, who has a severe level of dysarthria and highly unintelligible speech. The highest recognition accuracy of 78% was achieved when training the model on typical speech data for speaker DS02F.

The performance of the SI models vary among speakers with dysarthria. The difference is still observed even within the group of speakers with PD. Speakers DS02F and DS04F seem to share strong similarities with typical speech in their way of expressing emotions although this may not mean that these similarities are exactly the same for those two speakers. Although models that are trained on typical speech may not work for all speakers with dysarthria, it is, however, very helpful for those for whom it does work. That is because collecting sufficient amounts of emotional typical speech to train such a SI model, is generally easier than collecting SD data for a particular target dysarthric speaker. In addition, working with these larger typical data sets enables the use of deep learning techniques that may improve the model's performance.

From the four emotions recognition experiment, we find that 'neutral' utterances were mostly misrecognised as 'sad' utterances. This aligns with findings reported in the literature on typical speech, and is caused by the close positions of these two emotions in the arousal-valence space [35]. We also found from recognising pairs of emotions that anger/neutral and anger/sad are the easiest pairs to recognise. High accuracy results were achieved for most of the speakers, with an accuracy of 100% achieved for some of them. The recognition accuracy results are lower in the case of recognising four emotional categories compared to recognising two emotional categories but still above random. This is expected as when the number of emotional categories increases, the task becomes more challenging and the discrimination between classes becomes harder due to the close boundaries in the feature space between emotions.

Collecting dysarthric emotional data is a very challenging task. However, although a limited number of speakers with dysarthria is included in this study, promising results are found. We plan to work on improving the recognition results by investigating the performance of other feature sets and other classifiers, as well as training the model with only typical speakers close in age to the speaker with dysarthria. Also, we plan to investigate the effect of training the model with mixed emotional speech (typical and dysarthric speech) and see whether this would help in improving the results.

5. Acknowledgement

This research has been supported by the Saudi Ministry of Education, King Saud University, Saudi Arabia. The authors would like to thank the speakers whose participation made this study possible.

6. References

[1] J. R. Duffy, *Motor Speech disorders-E-Book: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2013.

[2] M. Walshe and N. Miller, "Living with acquired dysarthria: the speaker's perspective," *Disability and rehabilitation*, vol. 33, no. 3, 2011.

[3] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Interspeech*, 2012.

[4] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain, "Automatic selection of speakers for improved acoustic modelling: Recognition of disordered speech with sparse data," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 254–259.

[5] F. Xiong, J. Barker, and H. Christensen, "Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition," in *13th ITG-Symp*, 2018.

[6] J. Shor, D. Emanuel, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nollstadt *et al.*, "Personalizing asr for dysarthric and accented speech with limited data," *arXiv preprint arXiv:1907.13511*, 2019.

[7] J. Keshet, "Automatic speech recognition: A primer for speech-language pathology researchers," *International Journal of Speech-Language Pathology*, vol. 20, no. 6, 2018.

[8] K. Huang, C. Wu, Q. Hong, M. Su, and Y. Chen, "Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds," in *ICASSP*. IEEE, 2019.

[9] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *ICASSP*. IEEE, 2019.

[10] R. Patel, "Prosodic control in severe dysarthria: preserved ability to mark the question-statement contrast." *Journal of speech, language, and hearing research*, vol. 45, no. 5, 2002.

[11] ——, "Acoustic characteristics of the question-statement contrast in severe dysarthria due to cerebral palsy," *Journal of Speech, Language, and Hearing Research*, 2003.

[12] ——, "Phonatory control in adults with cerebral palsy and severe dysarthria," *Augmentative and Alternative Communication*, vol. 18, no. 1, 2002.

[13] L. Liu, M. Jian, and W. Gu, "Prosodic characteristics of mandarin declarative and interrogative utterances in parkinson's disease," in *Interspeech*, 2019.

[14] J. K.-Y. Ma and R. Hoffmann, "Acoustic analysis of intonation in parkinson's disease," in *Interspeech*, 2010.

[15] J. K.-Y. Ma, T. L. Whitehill, and S. Y.-S. So, "Intonation contrast in cantonese speakers with hypokinetic dysarthria associated with parkinson's disease," *Journal of Speech, Language, and Hearing Research*, 2010.

[16] G. J. Canter, "Speech characteristics of patients with parkinson's disease: I. intensity, pitch, and duration," *Journal of speech and hearing disorders*, vol. 28, no. 3, 1963.

[17] J. Illes, E. Metter, W. Hanson, and S. Iritani, "Language production in parkinson's disease: Acoustic and linguistic considerations," *Brain and language*, vol. 33, no. 1, 1988.

[18] A. Ghio, D. Robert, C. Grigoli, M. Mas, C. Delooze, C. Mercier, and F. Viallet, "F0 characteristics in parkinsonian speech: contrast between the effect of hypodopaminergy due to parkinson's disease and that of the therapeutic delivery of l-dopa," *Revue de laryngologie-otologie-rhinologie*, vol. 135, no. 2, 2014.

[19] V. L. Hammen and K. M. Yorkston, "Speech and pause characteristics following speech rate reduction in hypokinetic dysarthria," *Journal of communication disorders*, vol. 29, no. 6, 1996.

[20] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson's disease," *The journal of the Acoustical Society of America*, 2011.

[21] R. J. Holmes, J. M. Oates, D. J. Phyland, and A. J. Hughes, "Voice characteristics in the progression of parkinson's disease," *International Journal of Language & Communication Disorders*, vol. 35, no. 3, 2000.

[22] M. S. Hawley, S. P. Cunningham, P. D. Green, P. Enderby, R. Palmer, S. Sehgal, and P. O'Neill, "A voice-input voice-output communication aid for people with severe speech impairment," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 21, no. 1, 2012.

[23] F. Xiong, J. Barker, and H. Christensen, "Phonetic analysis of dysarthric speech tempo and applications to robust personalised dysarthric speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5836–5840.

[24] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: UK*, 2011, available at http://kahlan.eps.surrey.ac.uk/savee/.

[25] L. Alhinti, H. Christensen, and S. Cunningham, "An exploratory survey questionnaire to understand what emotions are important and difficult to communicate for people with dysarthria and their methodology of communicating," *International Journal of Psychological and Behavioral Sciences*, vol. 14, no. 7, pp. 187–191, 2020.

[26] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, 2016.

[27] L. Tian, J. Moore, and C. Lai, "Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features," in *IEEE Spoken Language Technology Workshop*, 2016.

[28] G. A. Trigeorgis, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP*. IEEE, 2016.

[29] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *arXiv preprint arXiv:1706.00612*, 2017.

[30] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, 2011.

[32] L. R. Brody and J. A. Hall, "Gender and emotion in context," *Handbook of emotions*, vol. 3, 2008.

[33] S. Yacoub, "Recognition of emotions in interactive voice response systems," in *EuroSpeech*, 2003.

[34] K. Dai, H. J. Fell, and J. MacAuslan, "Recognizing emotion in speech using neural networks," *Telehealth and Assistive Technologies*, vol. 31, 2008.

[35] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, 1980.