AUTOMATIC DETECTION OF EXPRESSED EMOTION IN PARKINSON'S DISEASE

Shunan Zhao[1], *Frank Rudzicz*[1,2], *Leonardo G. Carvalho*[1], *César Márquez-Chin*[2], *Steven Livingstone*[2,3]

[1] Department of Computer Science, University of Toronto; [2] Toronto Rehabilitation Institute;
[3]Department of Psychology; Ryerson University; Toronto Canada

ABSTRACT

Patients with Parkinsons Disease (PD) frequently exhibit deficits in the production of emotional speech. In this paper, we examine the classification of emotional speech in patients with PD and the classification of PD speech. Participants were recorded speaking short statements with different emotional prosody which were classified with three methods (naïve Bayes, random forests, and support vector machines) using 209 unique auditory features. Feature sets were reduced using simple statistical testing. We achieve accuracies of 65.5% and 73.33% on classifying between the emotions and between PD vs. control, respectively. These results may assist in the future development of automated early detection systems for diagnosing patients with PD.

***Index Terms*—** Parkinson's disease, emotion, classification, acoustic features

1. INTRODUCTION

Parkinson's disease (PD) is a sporadic neurodegenerative disease which primarily affects individuals of advanced age [1]. Its cardinal symptoms include akinesia (inability to initiate movement), tremor, rigidity, and postural imbalance [2, 3]. It is the most common neurodegenerative disorder after Alzheimer's disease [2]. In North America alone, there are approximately 1 million individuals with this disease [4] and there is currently no cure.

Early diagnosis of PD is critical to its treatment, which focuses on controlling symptoms and improving quality of life [2]. Unfortunately, diagnosing PD is challenging; current diagnosis methods are based on a patient's clinical history and through physical examination as there are currently no known biomarkers for diagnostic tests [1]. Appropriate management of PD requires regular monitoring of patients by a specialist, which can represent a significant burden on the health care system and on individuals themselves [2].

There has been increasing interest in creating tools that would improve the accuracy of PD diagnosis, support ongoing screening and monitoring, and allow for more rapid interventions [3]. One approach is to develop sophisticated speech analysis techniques to identify acoustic markers of the disease [5, 3, 2, 6, 5, 7]. Indeed, one of the earliest signs of PD consists of speech that may be softer, less distinct and with limited prosody (rhythm, stress, and intonation) [8].

The communication of emotion is an essential part of daily life. In PD, this capacity is often muted; patients exhibit deficits in the ability to produce and to respond to emotional tone of voice, facial expressions, and expressive body movements [9, 10, 11]. Patients with PD are often present with dysarthria, which is characterized by monotony of pitch and loudness, reduced stress, variable rate, imprecise consonants, and a breathy and harsh voice, all of which affect a patients ability to produce an emotional tone of voice [12].

1.1. Related work

Little et al. [13] classified between healthy adults and people with PD using a dataset of sustained phonations in which each of 31 speakers produce a single vowel at a constant pitch. They extracted various acoustic features and reported accuracies as high as 91%, although they provided very limited description of their method beyond the use of support vector machines (SVMs). Tsanas et al. [3] extended that work and performed feature selection on 132 acoustic features using 10-fold cross-validation with 100 repetitions. While they obtained similar accuracy, their method can apparently train the classifier using data from the same speakers used in classification. In practice, new subjects could not have been enrolled in training without first obtaining a diagnosis, so leave-one-out classification might have been more appropriate.

Previous work has not classified emotional speech in people with PD. In emotion detection, Forsell [14] showed that anger, despondency, and the level of emotional intensity can be determined from various acoustic features, such as the syllable rate, the mean of the pitch, and the first three formants.

2. EXPERIMENTAL SETUP

Five PD patients (mean age = 64.3, SD = 10.5, range = 24) and seven healthy age-matched adults (mean age = 62.6, SD = 7.4, range = 18) were recruited in Toronto. PD subjects had been diagnosed with idiopathic PD, were in Hoehn-and-Yahr stage 2, 2.5, or 3, had mild-to-no cognitive impairment, and mild-to-no clinical depression. All participants had normal or corrected-to-normal vision and hearing.

Cognitive functioning of PD patients was assessed with the Montreal Cognitive Assessment [15] (M = 27.0, SD = 1.7), where a score of ≥ 26 (out of 30) indicates no cognitive impairment. These results were in line with typical scoring for PD patients in Hoehn and Yahr stages 2 through 4 (Fahn, Elton, & Goldstein, 1987). Emotional status of PD patients was assessed with the Beck Depression Inventory [16] (M = 7.8, SD = 5.2), where a score of ≤ 16 is considered mild to no mood impairment.

All PD patients were currently taking part in a novel 13-week singing therapy designed to retrain deficits in facial motor function and vocal expressiveness [17]. The therapy uses an imitative facial and vocal mimicry paradigm, which focuses on the identification and reproduction of strong emotional displays, to retrain damaged motor and neural function.

2.1. Data collection

Participants were recorded individually using an Isomax EarSet E60P5L microphone. Participants were presented a randomly permuted set of 50 pre-recorded prompts consisting of emotional sentences spoken by a professional actress. The semantic content of each statement did not necessarily match its emotional prosody. For instance, the prompt "I've been crying all day" could be spoken in a happy tone. Prompts consisted of ten unique statements, each produced with one of the five prosodic emotions: anger, happy, sad, neutral, and fear. For each prompt, participants were told the prosodic emotion and instructed to repeat the prompt with that emotion, regardless of the prompt's semantic content. Recording sessions lasted approximately 25 minutes each.

2.2. Acoustic features

As in previous work [18], we measure pause-to-word ratio (i.e., the ratio of non-silent segments to silent segments longer than 150 ms), mean fundamental frequency (F0) and variance, total duration of speech, long and short pause counts ($>$ 0.4 ms and $>$ 0.15 ms, respectively) [19], mean pause duration, and phonation rate (the amount of the recording spent in voiced speech) [20]. We also include the mean and variance for the first 3 formants ($F1, F2, F3$), mean instantaneous power, mean and maximum first autocorrelation function, skewness, kurtosis, zero-crossing rate, and mean recurrence period density entropy (a method for measuring the periodicity of a signal, which has been applied to pathological speech generally [21]). Additionally, jitter [22] and shimmer are computed as:

$$\text{jitter}(x) = \frac{1}{N-1}\sum_{k=1}^{N-1} |P_0[k+1] - P_0[k]|$$

$$\text{shimmer}(x) = \frac{1}{N-1}\sum_{k=1}^{N-1} |x[k+1] - x[k]|$$

where $P_0(k)$ is the pitch period length ($1/F0$) at time k in a sequence x with N observations, and $x[k]$ is third-order median filtered. To this we also add the kurtosis and skewness of the 12^{th}-order autocorrelation linear predictive coding analysis of the signal, as well as the energy in the residual of this analysis (sometimes called the 'gain' of the filter). We also compute the variance, mean, kurtosis, and skewness for each of the 42 Mel-frequency cepstral coefficients, as well as the kurtosis and skewness of the means (i.e., instantaneous, δ, and $\delta\delta$ of each of the 13 first coefficients (0^{th} included) and log energy) taken over the entire utterance.

Aperiodicity and bradykinesia (slowed articulation) of speech are a core symptom of Parkinson's disease [23]. Therefore, our final subset of features involve recurrence quantification analysis (RQA) of the cross recurrence of each utterance. Specifically, for 3^{rd}-order RQA with delay 2, neighbourhood 0.5, we compute the mean recurrence rate, determinism, {mean, maximal, and entropy of} diagonal line length, laminarity, trapping time, maximal vertical line length, recurrence time of 1^{st} and 2^{nd} types, clustering coefficient, and transitivity over windows of length 1000 samples (with a 500-sample window shift).

2.2.1. Feature selection

Due to the high dimensionality ($d = 209$) of the feature space, we perform feature selection to reduce overfitting. For the PD vs. control classification task, we perform a t-test for each feature across the two groups. We then rank the features by p-value we use the n features with the lowest p-values. This can reveal discriminating features between the two groups [18]. For classifying between the five emotions, we perform an ANOVA test for every feature, where the factor is the emotion. As before, we rank the features by p-value and we use the n features with the lowest p-value, as in [24]. In both methods, $n \in \{10, 20, 40, \ldots, 200, 209\}$. Table 1 shows the 10 most discriminating features for each task.

2.3. Classification methods

We conduct four classification experiments, namely:

1. Leave-one-out, emotion. We test an emotion classifier with each speaker, trained with data from all others.
2. Leave-one-out, emotion, within group. Same as above, except training is done only on other speakers from the same group (PD or control) as the test subject.
3. Speaker-dependent, emotion. Each classifier is trained and tested with each speaker individually, using 10-fold cross-validation.
4. Leave-one-out, diagnostic. Same as 1., except classifiers are trained to differentiate PD from control.

We test three classifiers: naïve Bayes (NB), SVM, and random forests (RF). NB models the probability of the class label

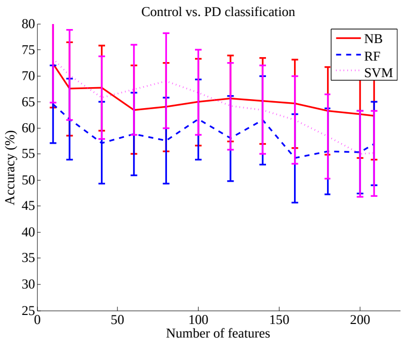

Fig. 1: Accuracies for the diagnostic task across classifiers. Bars correspond to 1 standard error from the mean.

given the features. It makes the simplifying assumption that the features are conditionally independent, given the class label. The SVM is a maximum-margin classifier capable of learning non-linear decision boundaries using a kernel function. RFs are ensemble classifiers consisting of multiple decision trees whose parameters are randomized; the output is the mode of the predictions of those trees.

Utterances were independently classified by two speech-language pathologists (SLPs). For each of the 600 utterances, SLPs identified both the participant group (PD or control) and the speaker's intended emotion (happy, angry, sad, fear, and neutral). Both SLPs had extensive training (> 12 months) and experience in working with individuals with PD, especially in outpatient/community settings.

3. RESULTS

The SLPs correctly identified PD (or its absence) 70.7% and 73.7% of the time from the voice alone, which is not significantly different from the best automated methods (73.3% on average), as shown in Figure 1.

The results for emotion classification are summarized in Figure 2 across classifiers and training methods. In general, the SVM outperforms the other classifiers, and these differences are always significant (e.g., a paired $t(22) = -11.83, p < 0.0001, CI = [-15.32, -10.75]$ for experiment 1 among PD test subjects). Using 140 features and performing leave-one-out cross validation over all speakers (both PD and controls), the SVM correctly identifies the spoken emotion 60% of the time, on average, among PD test subjects and of 69% of the time among controls. By comparison, human annotators are only accurate 39.4% of the time among PD test subjects (and 53.7% of the time among controls). An ANOVA across the different emotion classification experiments reveals that both the identity of the speaker and the classifier contribute significantly to the variance in accuracy ($p < 0.0001$).

	feature	p-value
Emotion	mean of the 5th MFCC coefficient	$1.39E^{-47}$
	mean F1	$2.82E^{-32}$
	ZCR	$4.71E^{-30}$
	mean of the 3rd MFCC coefficient	$1.60E^{-28}$
	28th CRQA coefficient	$1.90E^{-25}$
	skewness of the 5th MFCC coefficient	$9.73E^{-25}$
	mean first autocorrelation function	$1.07E^{-23}$
	9th CRQA coefficient	$1.13E^{-23}$
	variance of the 23rd MFCC coefficient	$3.47E^{-21}$
	variance of the 7th MFCC coefficient	$5.10E^{-20}$
PD vs. CTRL	mean instantaneous power	$1.83E^{-44}$
	mean of the 1st MFCC coefficient	$1.17E^{-37}$
	mean of the 2nd MFCC coefficient	$4.56E^{-35}$
	variance of the 4th MFCC coefficient	$6.99E^{-35}$
	mean RPDE	$6.54E^{-31}$
	kurtosis of the MFCC means	$1.25E^{-30}$
	fundamental frequency variance	$8.76E^{-28}$
	variance of the 18th MFCC coefficient	$1.19E^{-27}$
	skewness of the MFCC means	$2.91E^{-24}$
	11th CRQA coefficient	$6.37E^{-24}$

Table 1: Most discriminating features (with lowest 10 p-values) for each task.

The averages (over SLPs) of the confusion matrices for classifying emotion among control and PD test subjects are shown in Table 2, along with the equivalent matrices from the most accurate SVM trained with leave-one-out among both PD and control subjects. While confusion matrices do not vary significantly between SLPs, there are clear differences between human and automated performance, especially among PD test subjects. Interestingly, although each of the five emotions occur in equal measure, SLPs were hesitant to label speech as happy (10% among controls, 6% among PD) and were significantly more likely to label PD speech as 'neutral' (30.2% of cases) compared with control speech (13.4%), which fits with the etiology of the disorder.

CONTROL		Estimate				
		fear	sad	anger	happy	neutral
True	fear	*35*/**55**	*14*/**7**	*11.5*/**5**	*3.5*/**1**	*6*/**2**
	sad	*8*/**4**	*55.5*/**51**	*4*/**7**	*1*/**4**	*1.5*/**4**
	anger	*10*/**6**	*2.5*/**5**	*44.5*/**49**	*7*/**3**	*6*/**7**
	happy	*27.5*/**5**	*7.5*/**9**	*8*/**9**	*22.5*/**45**	*4.5*/**2**
	neutral	*3*/**8**	*23*/**9**	*14*/**6**	*1*/**4**	*29*/**43**

PD		Estimate				
		fear	sad	anger	happy	neutral
True	fear	*15.5*/**31**	*12*/**2**	*7*/**2**	*2*/**2**	*13.5*/**13**
	sad	*3.5*/**8**	*29*/**29**	*3.5*/**2**	*1.5*/**2**	*12.5*/**9**
	anger	*10*/**6**	*4.5*/**1**	*19.5*/**25**	*4*/**6**	*12*/**12**
	happy	*20.5*/**12**	*6.5*/**5**	*5*/**2**	*7.5*/**21**	*10.5*/**10**
	neutral	*1.5*/**4**	*18*/**0**	*3.5*/**1**	*0*/**1**	*27*/**44**

Table 2: Confusion matrices for identification of emotion. *Averages over SLPs are in italic* and **SVM values are in bold**.

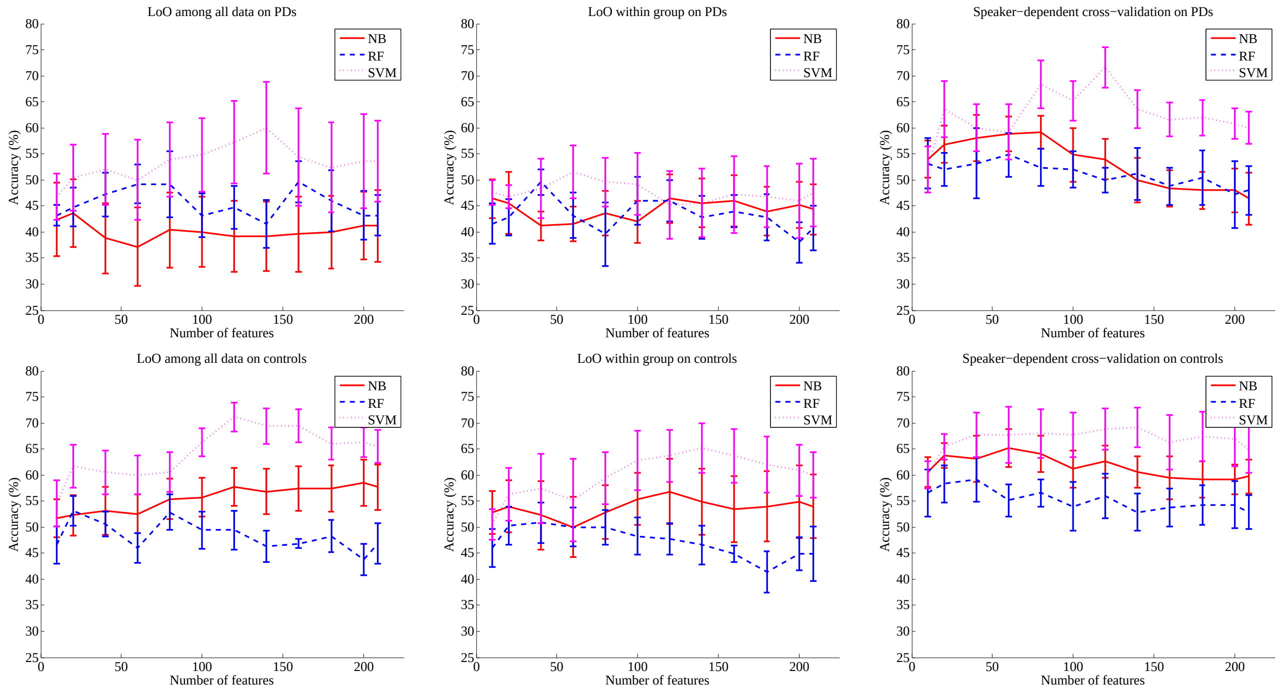

Fig. 2: Average accuracies for emotion classification across different experiments and classifiers. Error bars correspond to 1 standard error from the mean.

4. DISCUSSION

This paper presents the first automatic classification of emotions in the voice of patients with PD. For this task, and for identifying the presence of PD from the voice, MFCC features are important discriminators, as shown in Table 1, supporting similar previous work [3]. Even so, the number of features appears to have a much smaller effect on the classification accuracy than the classifier used.

Using the experimentation methodology in Tsanas et al. [3], where the classifier was trained on the speaker's own data, we are able to achieve much higher accuracies for both emotion classification (73.55%) and the PD vs. control task (83.78%). This approach would not be tenable in practice, however, since labeled data would not be available for previously unseen subjects.

Wilting et al. [25] conducted an experiment in which real emotions were induced from one group while another group was instructed to actively portray emotions. That work showed that acted emotions were perceived more strongly than natural emotions by human annotators. Since software tools to support therapy would ideally use imitated emotion rather than real emotion in the training of expressive speech, this type of data is appropriate.

We are continuing to record additional subjects within a 13-week therapy based on singing. In the future, we will examine the effects of that therapy on the ability of PD patients to effectively express their emotions. We will also incorporate visual features based on simultaneous 3D video recordings of the face during speech.

5. REFERENCES

[1] J. Meara and P. Hobson, “Epidemiology of Parkinson's disease,” in *Parkinson's Disease in the Older Patient*, pp. 30–38. Radcliffe Publishing, Jan. 2008.

[2] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, “Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests,” *Biomedical Engineering, IEEE Transactions on*, vol. 57, no. 4, pp. 884–893, 2010.

[3] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, “Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease,” *Biomedical Engineering, IEEE Transactions on*, vol. 59, no. 5, pp. 1264–1271, 2012.

[4] A. E. Lang and A. M. Lozano, “Parkinson's disease. First of two parts.,” *The New England journal of medicine*, vol. 339, no. 15, pp. 1044–1053, Oct. 1998.

[5] B. E. Sakar, M. M. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, “Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings,” *Biomedical and*

Health Informatics, IEEE Journal of, vol. 17, no. 4, pp. 828–834, 2013.

[6] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," *Journal of the Royal Society, Interface / the Royal Society*, vol. 8, no. 59, pp. 842–855, June 2011.

[7] H. Hazan, D. Hilu, L. Manevitz, L. O. Ramig, and S. Sapir, "Early diagnosis of Parkinson's disease via machine learning on speech data," *Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of*, pp. 1–4, 2012.

[8] G. Macphee, "Diagnosis and differential diagnosis of Parkinson's disease," in *Parkinson's Disease in the Older Patient*, pp. 41–75. Radcliffe Publishing, 2008.

[9] D. Bowers, K. Miller, W. Bosch, D. Gokcay, O. Pedraza, U. Springer, and M. Okun, "Faces of emotion in Parkinsons disease: micro-expressivity and bradykinesia during voluntary facial expressions," *Journal of the International Neuropsychological Society*, vol. 12, no. 6, pp. 765–773, 2006.

[10] K. Dujardin, S. Blairy, L. Defebvre, S. Duhem, Y. Noël, U. Hess, and A. Destée, "Deficits in decoding emotional facial expressions in Parkinson's disease," *Neuropsychologia*, vol. 42, no. 2, pp. 239–250, 2004.

[11] D. H. Jacobs, J. Shuren, D. Bowers, and K. M. Heilman, "Emotional facial imagery, perception, and expression in Parkinson's disease," *Neurology*, vol. 45, no. 9, pp. 1696–1702, 1995.

[12] S. Pinto, C. Ozsancak, E. Tripoliti, S. Thobois, P. Limousin-Dowsey, and P. Auzou, "Treatments for dysarthria in Parkinson's disease," *The Lancet Neurology*, vol. 3, no. 9, pp. 547–556, 2004.

[13] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *Biomedical Engineering, IEEE Transactions on*, vol. 56, no. 4, pp. 1015–1022, 2009.

[14] M. Forsell, "Acoustic correlates of perceived emotions in speech," M.S. thesis, Royal Institute of Technology, School of Computer Science and Communication, Stockholm Sweden, 2007.

[15] Viv Peto, C. Jenkinson, and R. Fitzpatrick, "PDQ-39: a review of the development, validation and application of a Parkinson's disease quality of life questionnaire and its associated measures," *Journal of Neurology*, vol. 245, no. 1, pp. S10–S14, 1998.

[16] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh, "An inventory for measuring depression," *Archives of general psychiatry*, vol. 4, no. 6, 1961.

[17] S. R. Livingstone, L. McGarry, P. van Lieshout, A. E. Lang, and F. A. Russo, "A novel singing therapy for rehabilitating facial and vocal expressive deficits in Parkinson's disease," Toronto, ON, 2013, Presented at the 9th Annual Toronto Rehabilitation Research Day.

[18] K. Fraser, F. Rudzicz, and E. Rochon, "Using text and acoustic features to diagnose progressive aphasia and its subtypes," in *Proceedings of Interspeech 2013*, Lyon France, August 2013.

[19] S. V. Pakhomov, G. E. Smith, D. Chacon, Y. Feliciano, N. Graff-Radford, R. Caselli, and D. S. Knopman, "Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration," *Cognitive and Behavioral Neurology*, vol. 23, pp. 165–177, 2010.

[20] B. Roark, M. Mitchell, J. Hosom, K. Hollingshead, and J. Kaye, "Spoken language derived measures for detecting mild cognitive impairment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2081–2090, 2011.

[21] M. A. Little, P. McSharry, I. Moroz, and S. Roberts, "Nonlinear, biophysically-informed speech pathology detection," in *Proceedings of ICASSP 2006*, Toulouse, France, 2006, pp. 1080–1083.

[22] D. Silva, L. Oliveira, and M. Andrea, "Jitter estimation algorithms for detection of pathological voices," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, 2009.

[23] S. C. Bhatnagar, *Neuroscience for the study of communication disorders*, Lippincott Wiliams & Wilkins, Baltimore Maryland, 2 edition, 2002.

[24] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[25] J. Wilting, E. Krahmer, and M. Swerts, "Real vs. acted emotional speech," in *Proceedings of Interspeech 2006*, 2006.