OBJECTIVE ASSESSMENT OF PATHOLOGICAL SPEECH USING DISTRIBUTION REGRESSION

Ming Tu[1], *Visar Berisha*[1,2], *Julie Liss* [1]

[1]Speech and Hearing Science Department
[2]School of Electrical, Computer, and Energy Engineering
Arizona State University

ABSTRACT

Objective assessment of pathological speech is an important part of existing systems for automatic diagnosis and treatment of various speech disorders. In this paper, we propose a new regression method for this application. Rather than treating speech samples from each speaker as individual data instances, we treat each speaker's data as a probability distribution. We propose a simple non-parametric learning method to make predictions for out-of-sample speakers based on a probability distance measure to the speakers in the training set. This is in contrast to traditional learning methods that rely on Euclidean distances between individual instances. We evaluate the method on two pathological speech data sets with promising results.

Index Terms— Distribution regression, divergence, objective assessment, speech pathology

1. INTRODUCTION

Clinical assessment in speech therapy is predominantly conducted through subjective tests performed by trained speech-language pathologists (SLPs). Subjective evaluations, however, can be inconsistent and unrepeatable [1], resulting in an inherent ambiguity about whether the patient is improving as a result of the therapy. One solution to this reliability problem has been the development of objective outcome measures that can automatically estimate how atypical a speech signal sounds to the average listener. In other words, these algorithms estimate an objective measure of the *perceived severity* of the speech disorder [2, 3, 4]. Existing objective assessment systems are usually based on regression analysis [3, 5, 6].

Objective outcome measures based on regression analysis require collecting speech samples from patients and having experts (e.g. SLPs) label the severity of each speaker on a subjective scale. Using this labeled data set, a data-driven model can be built to predict the intelligibility/severity of new speakers. A number of papers have appeared on this topic in the last few years. These studies have either focused on developing new and more sensitive features that measure various aspects of phonation, articulation, and prosody [7, 8]; or on developing more advanced machine learning algorithms to address the same problem[9, 10].

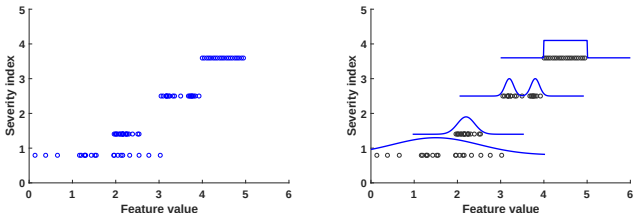

Fig. 1. A toy example shows the problem of existing methods and how we want to solve it.

However, there is one problem ignored by most existing data-driven regression models in this area. Evaluators usually rate speech at the speaker level rather than the sentence level, but there are often multiple stimuli produced by the same speaker. For example, in pathological speech, the data from each speaker often includes individual words, short sentences, and paragraphs. To fit this to the traditional single-instance learning paradigm, existing methods simply assign all sentences from the same speaker the same label, or they average the features from multiple sentences into a single instance [11]. This strategy may work well if the variation of each speaker's data is small; however this is rarely the case since there are often differences in recording conditions, spoken content, etc. from sentence to sentence that impact the distribution of the data.

We show a synthetic example in Fig. 1 that depicts this problem. Let us consider a single feature extracted from multiple sentences spoken by four speakers. Each of the four speakers is evaluated by an SLP as having a different severity index (value on the y-axis). We plot the multiple instances from each speaker (the one-dimensional feature along the x-axis) against the severity index (the value on the y-axis). If our aim is to predict the severity of the speaker based on the features, one can imagine multiple reasonable regression fit-

This research was supported in part by National Institute of Health, National Institute on Deafness and Other Communicative Disorders Grants Nos. 2R01DC006859 and 1R21DC012558.

s to the data, depending on the fitting criterion. We propose to develop criteria that consider the data from each speaker as a *distribution* rather than individual instances (see Fig. 1 right). By considering the data as samples drawn from an unknown distribution, we can develop cost functions based on distances between distributions rather than distances between individual instances.

Relation to previous work: In this paper, we propose a new method for objective assessment of pathological speech based on distribution regression. Instead of considering utterances from the same speaker as individual instances, we model the data from each speaker as a probability density function (PDF) and consider the multiple instances as samples drawn from that PDF. This approach has been explored in multiple instance classification [12, 13]; here we extend it to regression and evaluate it on an application where we predict the severity of a dysarthric speaker directly from the speech samples.

We evaluate the algorithm on the Parkinson's condition database from the 2015 INTERSPEECH Computational Paralinguistics Challenge and a dysarthric speech corpus collected in our lab. We observe that the method is particularly useful for cases where there exist mismatched recording conditions between the training corpus and the test corpus.

2. PROPOSED ALGORITHM

We consider a database of N_{s} speakers with n speech samples from speaker k, $k \in [1 \dots N_{\mathrm{s}}]$. We assume that we extract a set of features from every speech sample from every speaker and denote the resulting feature vector by $\mathbf{x}_{k_i}$, $i \in [1 \dots n]$ (i^{th} sample of speaker k). In multiple instance regression, speaker k has a set of feature vectors $S_k = \{\mathbf{x}_{k_1}, \mathbf{x}_{k_2}, ..., \mathbf{x}_{k_n}\}$ and only one label Y_k. We denote the labeled training data by $\mathcal{S} = \{(S_1, Y_1), (S_2, Y_2), \dots, (S_{N_{\mathrm{s}}}, Y_{N_{\mathrm{s}}})\}$. In examples involving speech, the multiple feature vectors could represent different sentences from the same speaker, different recording conditions, or other sources of variability. Our goal is to find a mapping function f that minimizes the error between $f(S_k)$ and Y_k:

$$f^* = \arg\min_f \sum_{k=1}^{N_{\mathrm{s}}} \mathcal{D}(f(S_k), Y_k), \tag{1}$$

where $\mathcal{D}$ is some distance measure between the predicted label and actual label. A common instantiation of this framework is the traditional single instance learning paradigm where the same label is copied to every sentence of the same speaker and the ℓ_2-error is used to measure the difference between the predicted label and the actual label,

$$f^* = \arg\min_f \sum_{k=1}^{N_{\mathrm{s}}} \sum_{i=1}^{n} (f(\mathbf{x}_{k_i}) - Y_k)^2. \tag{2}$$

Rather than adopting a specific hypothesis class (e.g. f is the class of linear functions), a non-parametric class of estimators based on nearest neighbors can be used [14]. An example of this is k-nearest neighbor (KNN) regression, where we use the training data $\mathcal{S}$ to make out-of-sample predictions about a test sample, S_{t}. Traditional KNN regression for single instance learning simply finds the K nearest neighbors of the test sample in the training set and computes the average of the labels,

$$f(S_t) = \frac{1}{K} \sum_{i \in NN(S_t, S)} Y_i, \tag{3}$$

where $NN(S_t, \mathcal{S})$ is the set of all nearest neighbors of S_t in $\mathcal{S}$.

For single instance learning, these nearest neighbors can be estimated using Euclidean distance measures; however in our method, each data point consists of a *distribution*. As a result, we propose to treat the acoustic features $\{\mathbf{x}_{k_1}, \mathbf{x}_{k_2}, ..., \mathbf{x}_{k_n}\}$ extracted from sentences produced by speaker k as samples drawn from this *unknown* distribution. To that end, we want to estimate distances between distributions rather than individual instances. Next, we introduce three metrics for measuring distance/divergence between two distributions.

2.1. Hausdorff distance

The Hausdorff distance measures how far two subsets of a metric space are from each other. It directly gives the distance between two sets of samples,

$$D_H(X, Y) = \max\{\sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y)\}, \tag{4}$$

where $d(x, y)$ is some distance measure between instances with the Euclidean distance commonly used. The Hausdorff distance has been used in both multiple instance learning (MIL) and multi-instance multi-label learning for classification [12, 15]. This measure is a simple extension of a single-instance metric to multiple instances and not a true divergence measure between two distributions. The next two measurements directly calculate a divergence between the data distributions.

2.2. Rényi-α divergence

The Rényi-α divergence is a family of divergence measures between distributions and is defined as

$$R_\alpha(p, q) = \frac{1}{\alpha - 1} \log \int p^\alpha(x) q^{1-\alpha}(x) dx, \tag{5}$$

where p and q are two distributions and α can be modified to achieve different divergences such as the Bhattacharyya distance when $\alpha = \frac{1}{2}$ and the Kullback-Leibler (KL) divergence when $\alpha \to 1$. In [13], the authors proposed a non-parametric estimator for the Rényi-α divergence based on density estimation and applied this estimator to several machine learning algorithms. In our example, p and q model the distributions of the features extracted from two speakers.

2.3. D_p divergence

The D_p divergence is a recently proposed divergence measure between distributions that can be estimated directly from samples drawn from those distributions without requiring parametric assumptions [16]. It is defined as

$$D_\alpha(p,q) = \frac{1}{4\alpha(1-\alpha)}\left[\int \frac{(\alpha p(x)-(1-\alpha)q(x))^2}{\alpha p(x)+(1-\alpha)q(x)}dx - (2\alpha-1)^2\right], \quad (6)$$

where p and q are two distributions and α is the prior probability of p. This divergence is guaranteed to provide a tighter bound on the Bayes classification error rate than Bhattacharyya distance and has been applied to several applications related to statistical learning [17, 18, 19]. This divergence can also be estimated non-parametrically without estimation or plug-in of the densities p and q.

In our proposed method, we evaluate these three measures as distances between speaker distributions using a nearest neighbor regression rule. We predict the label of a new unseen test speaker by using eqn. (3). For the Rényi-α divergence, we set $\alpha = 0.99$ to approximate the KL divergence. For the D_p divergence we set α to 0.5.

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1. Feature description

Before feature extraction, speech samples are first downsampled to 16kHz. The extracted speech features include five subsets: The envelope modulation spectrum [20], a representation of the slow amplitude modulations in a signal; the long-term average spectrum features [21] and the Mel-Frequency Cepstral Coefficients (MFCC) statistics capture atypical average spectral information in the signal; dysphonia features capture a patients' ability to control glottal movement; correlation structure features [22, 8] capture the evolution of the vocal tract shape and dynamics at different time scales via auto- and cross- correlation analysis of formant tracks and MFCCs. The total number of features is 1201.

3.2. INTERSPEECH 2015 Computational Paralinguistics Challenge dataset

This data set consists of speech samples from 61 Spanish-speaking patients (30 females) with Parkinson's disease (42 speech samples per person). There are 35 patients in the training set, 15 in the development set and 11 in the test set. Speakers in the test set were recorded in a different environment than the training and development set. As a result, the test audio samples were noticeably degraded by background noise (people talking, outside traffic noise, etc), whereas the training and development speech samples were relatively clean. Each speaker is evaluated by clinicians based on the Unified Parkinson's Disease Rating Scale (UPDRS) (one label per speaker). The goal of the Parkinson's condition challenge is to predict the UPDRS score directly from the speech samples. More detailed information on the data and the challenge can be found in [6]. Since some of the utterances were very short, we are restricted to only the first two scales of the correlation structure features described in [22]. Thus, for this data set, the feature dimension is reduced from 1201 to 809. We calculate the nearest neighbors for each test speaker using the citation KNN rule which not only counts the closest K_1-neighbors of S_t but also the speakers that consider S_t as a top-K_2-closest neighbor [12]. The nearest neighbors are then used to estimate the label of the speaker using eqn. (3). This is done using the three different distribution distance metrics described in the previous section and we vary K_1 (from 1-5) and K_2 (from 0-5) to find optimal performance.

Table 1. Performance comparison with baseline system.

	Baseline	Hausdorff	D_p	Rényi_KL
CV	0.270	0.444	0.266	**0.499**
Development	0.699	0.337	**0.746**	0.729
Test	0.298	0.326	0.496	**0.498**
Test Optimal	0.357	**0.578**	0.501	0.552

We measure the performance of the algorithm using the Pearson correlation coefficient between the predicted UPDRS labels and the true UPDRS labels at the speaker level. We compare our approach (using the three different distances in Section 2) against that of the baseline system that uses single-instance learning based on support vector regression (SVR) [6]. We select the parameter C for the SVR algorithm from among $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ to find the optimal performance. For the single instance implementation, we average the predicted UPDRS of all sentences produced by one speaker and use that as the predicted UPDRS of that speaker.

In Table 1, we show the results from four evaluation conditions: "CV: 7-fold cross-validation performance on the training set", "Development: performance on the development set" and "Test: performance on the test set with parameters set using the training and development set" and "Test Optimal: performance on the test set with parameters that yield the highest performance". The table compares the performance of the baseline system and our proposed method using the three different distance measurements between distributions. "Hausdorff" represents Hausdorff distance, "D_p" represents D_p divergence and "Rényi_KL" represents the approximation of the KL divergence. The results show that in all cases the nearest neighbor approach based on distribution distances provide considerable improvement compared to the baseline method.

3.3. ASU Dysarthric speech database

The data, collected in the Motor Speech Disorders Lab at ASU, consists speech samples from 56 dysarthric patients (native English speakers) with four different dysarthria subtype-

Table 2. Comparison of Pearson correlation coefficients among five different methods.

	wo_PCA				w_PCA			
	clean	Noise	Reverb	Noise+Reverb	clean	Noise	Reverb	Noise+Reverb
LR	0.563	0.357	0.243	0.037	**0.782**	0.596	0.488	0.153
SVR	0.728	0.500	0.477	0.012	0.759	0.600	0.453	0.243
Hausdorff	0.530	0.477	0.580	0.446	0.589	**0.617**	0.503	0.497
DP	0.737	0.578	0.660	0.326	0.773	0.581	0.648	0.363
Renyi_KL	**0.777**	**0.612**	**0.677**	**0.550**	0.766	0.615	**0.698**	**0.660**

s: ataxic dysarthria secondary to cerebellar degeneration (n = 15), mixed flaccid-spastic dysarthria secondary to amyotrophic lateral sclerosis (n = 15), hyperkinetic dysarthria secondary to Huntington's disease (n = 6), and hypokinetic dysarthria secondary to Parkinson's disease (n = 20). Speech materials included the standard "Grandfather" paragraph and 5 rhythm sentences [23, 24]. The "Grandfather" paragraph recordings were segmented into individual sentences. This results in a total of 40 sentences for each speaker. We asked 15 master's students from the ASU SLP program to rate the severity of each patient based on their produced speech on a 1-7 (typical-severely atypical) scale. To integrate ratings by multiple raters, we split the 15 raters into two groups - one set is used to train the model (7), the other set is used to test the model (8). For each of the two groups, we use the Evaluator Weighted Estimator (EWE) [25] to combine the multiple ratings into a single set of ratings by calculating the mean value weighted by individual reliability.

To evaluate the proposed algorithm, we use 5-fold cross-validation. For each fold, 45 speakers are used for training and the remaining 11 speakers are used for validation. After obtaining the predicted severity ratings of all speakers, we calculate the Pearson correlation coefficient between the EWE rating from the test group raters and the predicted ratings. We evaluate the three distance measurements and compare our method with two commonly used single-instance based methods: linear regression (LR) and SVR. For training the single-instance algorithms, each sentence is assigned the same label (that of the speaker) and the predicted rating of a new speaker is the average predicted rating of all sentences produced by that speaker. We search the same parameter space as in the previous section (K_1, K_2 for the citation KNN rule and the parameter C for the SVR algorithm) as in the previous experiment.

In the second part of the experiment, we evaluate the robustness of our method to mismatched recording conditions between the training and test data. For all experiments here, we use clean data to train and noisy data to evaluate. We design three conditions by using 12 types of reverberation (room impulse responses from the REVERB challenge 2014 [26]) and 4 types of noise (babble, computer keyboard, eating chips and ambient noise). For the first condition ("Noise" in table 2) we randomly select one type of background noise for the "Grandfather" passage sentences and another type of noise for the 5 rhythm sentences. The signal-to-noise ratio (SNR) is set at 15dB. In the second condition ("Reverb" in table 2) we randomly select one type of reverberation for the "Grandfather" passage sentences and another type of reverberation for the 5 rhythm sentences. In the third condition ("Noise+Reverb" in table 2) we randomly select one type of noise (15dB SNR) and one type of reverberation for the "Grandfather" passage sentences and another type of noise and another type of reverberation for the 5 rhythm sentences.

For all experiment conditions, we test two scenarios: one without principal components analysis (PCA) after feature extraction ("wo_PCA" in table 2) and one with PCA to reduce the dimension to 100 ("w_PCA" in table 2). The cross-validation results are shown in Table 2 (the correlation coefficients represent optimal performance). The table shows that for the majority of cases our proposed distribution regression method outperforms the single-instance methods. Though PCA can reduce the impact of additive noise to some extent, distribution regression based methods are consistently more robust to both reverberation and additive noise.

4. CONCLUSIONS

In this paper we propose a new method to tackle the problem of objective assessment of pathological speech. Instead of single-instance learning, we treat the training data from each speaker as samples from a distribution. A simple and efficient lazy learning model KNN is used to make predictions on out-of-sample speakers based on their distance to speakers in the training set. We apply three distance measurements between distributions and compare our method with baseline single-instance based methods. Experiments on two data sets show the advantage of our method over existing single-instance based methods, especially for the case of mismatched train and test conditions. While the focus of this work was pathological speech analysis, the approach can be applied to other computational paralinguistic tasks.

5. ACKNOWLEDGEMENTS

We would like to thank Dr. Juan Rafael Orozco for providing the Parkinson's condition data set of the INTERSPEECH 2015 Computational Paralinguistics Challenge.

6. REFERENCES

[1] Kate Bunton, Raymond D Kent, Joseph R Duffy, John C Rosenbek, and Jane F Kent, "Listener agreement for auditory-perceptual ratings of dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 6, pp. 1481–1495, 2007.

[2] Visar Berisha, Julie Liss, Steven Sandoval, Rene Utianski, and Andreas Spanias, "Modeling pathological speech perception from data with similarity labels," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 915–919.

[3] Catherine Middag, Jean-Pierre Martens, Gwen Van Nuffelen, and Marc De Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 1–9, 2009.

[4] Andreas Maier, Tino Haderlein, Florian Stelzle, Elmar Nöth, Emeka Nkenke, Frank Rosanowski, Anne Schützenberger, and Maria Schuster, "Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1, 2010.

[5] Dávid Sztahó, Gábor Kiss, and Klára Vicsi, "Estimating the severity of parkinson's disease from speech using linear regression and database partitioning," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] Björn Schuller, Stefan Steidl, Anton Batliner, Simone Hantke, Florian Hönig, Juan Rafael Orozco-Arroyave, Elmar Nöth, Yue Zhang, and Felix Weninger, "The interspeech 2015 computational paralinguistics challenge: Nativeness, parkinsons & eating condition," in *Proceedings of Interspeech*, 2015.

[7] Guozhen An, David Guy Brizan, Min Ma, Michelle Morales, Ali Raza Syed, and Andrew Rosenberg, "Automatic recognition of unified parkinsons disease rating from speech with acoustic, i-vector and phonotactic features," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] James R Williamson, Thomas F Quatieri, Brian S Helfer, Joseph Perricone, Satrajit S Ghosh, Gregory Ciccarelli, and Daryush D Mehta, "Segment-dependent dynamics in predicting parkinsons disease," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] Dingchao Lu and Fei Sha, "Predicting likability of speakers with gaussian processes," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[10] Alexander Zlotnik, Juan M Montero, Rubén San-Segundo, and Ascensión Gallardo-Antolín, "Random forest-based prediction of parkinson's disease progression using acoustic, asr and intelligibility features," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] Zhuang Wang, Liang Lan, and Slobodan Vucetic, "Mixture model for multiple instance regression and applications in remote sensing," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 50, no. 6, pp. 2226–2237, 2012.

[12] Jun Wang, "Solving the multiple-instance problem: A lazy learning approach," in *In Proc. 17th International Conf. on Machine Learning*, 2000.

[13] Barnabás Póczos, Liang Xiong, and Jeff Schneider, "Nonparametric divergence estimation with applications to machine learning on distributions," *arXiv preprint arXiv:1202.3758*, 2012.

[14] Thomas Cover and Peter Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[15] Zhi-Hua Zhou, Min-Ling Zhang, Sheng-Jun Huang, and Yu-Feng Li, "Multi-instance multi-label learning," *Artificial Intelligence*, vol. 176, no. 1, pp. 2291–2320, 2012.

[16] Visar Berisha, Alan Wisler, Alfred O Hero III, and Andreas Spanias, "Empirically estimable classification bounds based on a nonparametric divergence measure," *Signal Processing, IEEE Transactions on*, vol. 64, no. 3, pp. 580–591, 2016.

[17] Visar Berisha and Alfred O Hero, "Empirical non-parametric estimation of the fisher information," *Signal Processing Letters, IEEE*, vol. 22, no. 7, pp. 988–992, 2015.

[18] Ming Tu, Visar Berisha, Martin Wolf, Jaesun Seo, and Yu Cao, "Ranking the parameters of deep neural networks using the fisher information," *ICASSP*, 2016.

[19] Alan Wisler, Visar Berisha, Julie Liss, and Andreas Spanias, "Domain invariant speech features using a new divergence measure," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 77–82.

[20] Julie M Liss, Sue LeGendre, and Andrew J Lotto, "Discriminating dysarthria type from envelope modulation spectra," *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 5, pp. 1246–1255, 2010.

[21] Phil Rose, *Forensic speaker identification*, CRC Press, 2003.

[22] James R Williamson, Thomas F Quatieri, Brian S Helfer, Rachelle Horwitz, Bea Yu, and Daryush D Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 41–48.

[23] Joseph R Duffy, *Motor speech disorders: Substrates, differential diagnosis, and management*, Elsevier Health Sciences, 2013.

[24] Julie M Liss, Laurence White, Sven L Mattys, Kaitlin Lansford, Andrew J Lotto, Stephanie M Spitzer, and John N Caviness, "Quantifying speech rhythm abnormalities in the dysarthrias," *Journal of Speech, Language, and Hearing Research*, vol. 52, no. 5, pp. 1334–1352, 2009.

[25] Michael Grimm, Kristian Kroschel, Emily Mower, and Shrikanth Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10, pp. 787–800, 2007.

[26] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, E-manuël AP Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, B-hiksha Raj, et al., "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–19, 2016.