



# Automatic estimation of Parkinson’s disease severity from diverse speech tasks

Jangwon Kim<sup>1</sup>, Md Nasir<sup>1</sup>, Rahul Gupta<sup>1</sup>, Maarten Van Segbroeck<sup>1</sup>, Daniel Bone<sup>1</sup>, Matthew Black<sup>1,2</sup>  
 Zisis Iason Skordilis<sup>1</sup>, Zhaojun Yang<sup>1</sup>, Panayiotis Georgiou<sup>1</sup> and Shrikanth Narayanan<sup>1</sup>

<sup>1</sup>Signal Analysis & Interpretation Laboratory, Univ. of Southern California, Los Angeles, CA, USA

<sup>2</sup>Information Sciences Institute, Univ. of Southern California, Marina del Rey, CA, USA

<sup>1</sup><http://sail.usc.edu>, <sup>2</sup>[www.isi.edu](http://www.isi.edu)

## Abstract

The need for reliable, scalable and efficient diagnosis of Parkinson’s Disease (PD) is a major clinical need. Automating the diagnosis can lead to more accurate and objective predictions as well as provide insights regarding the nature of Parkinson’s condition. This paper proposes a fully automated system to rate the severity (UPDRS-III scale) of PD from patients’ speech. Specifically, the system captures atypicalities in an individual’s voice when performing multiple diverse speaking tasks and makes a unified prediction of the PD severity. The performance is tested in a cross-data setting, with different subjects and dissimilar recording conditions. Results indicate that (i) effective features vary depending on the nature of the specific speech task, (ii) additional novel feature sets to detect distortions in Parkinson’s speech significantly improve the prediction accuracy from the Interspeech15 Challenge baseline system and (iii) our fusion system based on an unsupervised clustering technique also improves the accuracy. Our system incorporates i-vector and functionals for segmental features, non-linear time series features, speech rhythm and automatic speech recognition decoding based features. By its application on the Interspeech15 eating condition challenge, the system also shows its potential for detecting other sources of speech variability.

**Index Terms:** Pathological speech, Automatic severity estimation, Parkinson’s disease

## 1. Introduction

Parkinson’s Disease (PD) is a neuro-degenerative disorder affecting the quality of life profoundly. Degeneration of dopamine-producing cells (dopaminergic neurons) in the brain causes defects of speech motor controls, resulting in a variety of atypicalities in the produced speech sound. Although patient-dependent, common effects of PD on the speech signal include reduced loudness (hypophonia), reduced pitch inflection (hypoprosodia), reduced stress, breathy and hoarse voice quality (dysphonia), imprecise articulation, defective speech rate, and rhythm [1, 2, 3, 4]. Various types of speech disorders motivate *holistic* information processing on pathological speech: capturing broad spectrum of speech acoustic cues. Over the last few decades, there has been sporadic interest in the characterization of those symptoms as well as the assessment of severity from paralinguistic cues in speech of Parkinson’s patients.

Evaluation of PD severity is essential for constant therapy and monitoring of the PD patients. Despite the huge demand for objective, accurate and robust assessment in clinical practice, the state-of-the-art evaluation method still relies on subjective judgments by human experts, which are costly and time-consuming. In this regard, there have been efforts to develop an

automatic evaluation system, especially using the speech of the patients. The use of speech has advantages over other diagnosis methodologies, because (i) most of the PD patients suffer from speech disorders and (ii) data acquisition is relatively easy and convenient, and can be done remotely, and continuously.

In the literature, most early studies [5, 6] are based on specific speech tasks, such as sustained vowels or specific word lists. This is useful for minimizing acoustic variability from other sources, e.g., lexical variations, and facilitates robust feature extraction. However, such a limited speech task is not ideal for the patients to display their various types and facets of speech disorders, e.g., in intonation, rhythm, articulation and breath control. Recently, there have been efforts toward developing a *comprehensive* analysis which considers both segmental and supra-segmental aspects of speech production, based on diverse speech tasks, including read speech tasks of various lengths, spontaneous speech and fast repeating speech (each exercising different facets of the production system). The goal of the automatic evaluation evolved to clinically more relevant: from binary decisions [7, 8] (the presence of PD or not) to regression in realistic assessment scales, e.g., the Unified Parkinson’s Disease Rating Scale (UPDRS) [2]. The present paper proposes a *fully* automated method to rate the severity of PD patients using vocal data drawn from diverse speech tasks with the audio recordings collected under the Interspeech 2015 Parkinson’s Condition (PC) Sub-Challenge [9].

The novelty of the proposed method lies in both feature engineering and the applied machine learning methods in the system development. First, we found that non-linear time series analysis on (estimated) glottal source signal and its delta are useful for rating the PD severity for various speech tasks. Second, we found that multi-level unsupervised clustering schemes help to provide a more accurate clustering for similar UPDRS scores. Finally, our system provides, in a fully automatic manner, a joint rating score for speech of Parkinson’s patients collected from diverse speaking tasks.

In addition to the PC Sub-Challenge, we also target in this paper the Eating Condition (EC) Sub-Challenge. The idea is to explore how the feature engineering and machine learning can be repurposed for an entirely different domain. For the EC task, we apply our system on speech recorded from different speakers with the goal to classify the eating conditions, i.e. presence or absence of food and type of food, when they were speaking [9].

The outline of the paper is as follows. We explain the details of the data pre-processing and unsupervised clustering methods in Sections 2 and 3. Sections 4 and 5 describe our feature sets, regressors and fusion systems. Experimental results and discussion are provided in Section 6. Finally, we conclude in Section 7 with a summary and future directions.

## 2. Data and Pre-processing

See [9, 10] for the details of the datasets used for the PC Sub-Challenge and the EC Sub-Challenge. The baseline feature sets, namely “ComPaRE set” comprise functionals of low-level descriptor (LLD), i.e., prosodic, spectral, voice quality and voice source features. The feature sets extracted using the latest version (ver. 2.1) of OpenSMILE [11] are provided.

To predict transcripts (phone sequences) of the dataset, we performed Automatic Speech Recognition (ASR) using KALDI [12]. A triphone acoustic model was trained on Mexican Spanish Broadcast news corpus [13] to learn the characteristics of normal Spanish speakers’ speech. It is noted that the performance of the present system can be improved by using speech data of Colombian Spanish for model training. Two language models were used; one was a unigram model trained using only the stimuli for the read speech utterances, while the other was a trigram model trained using the entire Mexican Spanish corpus only for monologue. Phone sequences of each utterance were decoded from corresponding phone lattice.

In order to minimize the variability due to speech tasks (task information was not provided in the test set), we estimated the type of speech task of each utterance, based on the phone sequences. For the training and development sets, the speech task label for each utterance is assigned depending on corresponding stimulus as follows: ‘0’ for isolated words, ‘1’ for rapid repetition of syllables, ‘2’ for text and monologue (long speech), and ‘3’ for sentences. We used a  $k$ -Nearest-Neighbor (KNN) classifier with Euclidean distance metric to estimate the task label (0–3) of each utterance. The counts for individual phones were used as the features for the classifier. Results of four-folds cross-validation on the training set indicates that  $k=1$  shows the best performance: 98.9%, 97.1%, 98.2% and 97.9% for each fold, respectively. KNN classification accuracy on the development set with  $k=1$  is 98.9%, which indicates reliable estimation performance of this method.

In order to remove silence and short pauses in the speech waveforms, we performed Voice Activity Detection (VAD) based on Root Mean Squared (RMS) energy; Voiced if RMS energy of the frame  $> 10\%$  of the 0.9 quantile in the utterance; Unvoiced otherwise.

## 3. Unsupervised Clustering

Our working hypothesis for using unsupervised clustering is that predicting UPDRS labels *jointly* for utterances that are very close in the acoustic feature space can reduce rating error. In fact, a joint rating approach makes sense for the PC Sub-Challenge, because the level of PD severity is evaluated for individual patients, not for each speech utterance. We tested this hypothesis using a single Gaussian-based bottom-up agglomerative hierarchical clustering method [14]. Linear predictive coding was used to represent the acoustic feature space. Generalized likelihood ratio [15] was used as inter-cluster distance measure. This off-line clustering method has shown satisfactory performance improvement for predicting other speaker traits in previous studies [16, 17].

Clustering accuracy depends on the spoken content, because the acoustic characteristics of speech from the same stimulus can be similar. Hence, we performed multi-level clustering in order to minimize clustering error due to lexical similarity. First, we performed utterance-level clustering within (estimated) speech task. Next, we determine the optimal pairs of clusters across tasks in order to apply task-dependent cluster

	Measure	Mean	STD
All-task clustering	Majority ratio	0.70	0.24
	STD of labels	8.19	5.14
Within-task clustering	Majority ratio	0.90	0.18
	STD of labels	3.08	5.35
Multi-level clustering	Majority ratio	0.85	0.11
	STD of labels	6.35	3.86

Table 1: Clustering performance on the development set in terms of (i) majority ratio and (ii) standard deviation (STD) of UPDRS labels. ‘Original’ indicates clustering on all speech tasks together (baseline); ‘Within-task clustering’ indicates the bottom level clustering; ‘Multi-level clustering’ indicates the merged clusters.

merging. For the distance metric between clusters, we examined metric functions of derived from utterance-wise distances (mean, minimum, maximum and quantiles). For example, the metric min. indicates the Euclidean distance of the pair of the closest files, one from each cluster (the two clusters are from different speech tasks). We compared their performances in terms of the mean of the difference of UPDRS labels in the final (merged) clusters. Table 1 shows that the multi-level clustering performs better than all-task clustering in terms of both majority ratio and standard deviation of labels, and close to within-task clustering which is the upper bound method. These clustering outputs are used for joint rating techniques, which will be described in Section 5.

## 4. Feature extraction

This section discusses a variety of speech features to capture atypicalities in PD patients’ speech, displayed in diverse speech tasks. Frame-level features are transformed to utterance-level features using (i) statistical functionals and (ii) i-vector system approach. Other utterance-level features (e.g., stability and irregularity in time series) are also examined. Details of the frame-level and utterance-level features are provided in below.

### 4.1. Frame-level features

Our spectral features comprise Mel-Frequency Cepstral Coefficients (MFCCs), Mel-Frequency Banks (MFBs), spectral shape functionals features ([10.0 25.0, 50.0, 75.0, 90.0] rolloffs, flux, entropy, variance, skewness, kurtosis, slope and their derivatives), Gammatone Frequency Cepstral Coefficients (GFCCs) [18], Gabor features (GBF) [19], spectro-temporal modulations [20], and long-term spectral variability profile [20]. Prosodic features consist of  $f_0$  and Root-Mean-Squared (RMS) energy. Voice quality features comprise Harmonics-to-Noise-Ratio (HNR), jitter and shimmer. We used Praat [21] to extract HNR, and OpenSMILE [11] to extract MFCCs, MFBs, spectral shape features, jitter and shimmer, with 25 msec window and 10 msec window shifting.

Phone posteriors computed from the ASR lattice can reflect acoustic variability. Specifically, a spiky posterior distribution may indicate high confidence of the normal speech model [22], implying that the spoken utterance matches the acoustic characteristics of normal speakers well. The posterior distribution is characterized using an entropy measure: lower (higher) entropy indicates more (less) spiky distribution. The triphone acoustic model in Section 2 is used for generating posteriors (one for each phone class) for each time frame in an utterance. Then, we compute entropy from the posteriors for each time frame.

## 4.2. Utterance-level features

The frame-level feature streams of the previous section are transformed into utterance-level features. To this end, we examine the use of (i) functionals and (ii) an i-vector extraction methodology.

The functionals comprises [0.1 0.25 0.5 0.75 0.9] quantiles, interquartile range, kurtosis and skewness of the feature streams, and were applied on the speech-only regions of the audio recordings. We also computed the functionals in the consonant regions and vowel regions separately.

In the i-vector system, we first trained a Universal Background Model (UBM) on each feature representation individually using all available training and development set data. To increase the noise robustness, mean- and variance-normalization on a per utterance basis was applied on all feature streams. We then trained a single i-vector subspace by jointly exploiting all UBMs using the UBM-fused total variability modeling technique that was proposed in [23]. The resulting i-vectors then yield a low dimensional utterance-level representation on which subsequent regressors can be applied. On the EC task, we found that an additional speaker normalization applied on the i-vector space further increases the classification accuracy. We refer to our paper on the Interspeech 2014 Challenge [24] for more details on the i-vector system and the speaker normalization strategy.

### 4.2.1. Non-linear time series analysis

One of our working hypotheses is that aperiodicity and irregularity exhibited in pathological speech is correlated with the severity of Parkinson’s disease. In the literature, prior work has pointed to the promise of Nonlinear Time Series Analysis (NTSA) in capturing PD related atypicalities in the speech waveform of sustained vowels [6]. In the present paper, we explore the utility of a few NTSA methods to capture the (nonlinear) atypicalities in the prosodic feature streams (f0, RMS energy after smoothing), glottal source signal and its delta, and the speech waveform for the four speech tasks. The glottal source signal was estimated using the pitch synchronous iterative adaptive inverse filtering method [25].

The list of NTSA methods we examined is as follows:

- Correlation Dimension (CD): This dimension provides cues of the geometric complexity of the time series in the phase space. Higher dimension indicates more complex dynamics of the signal. The present study determines the optimal CD based on Taken’s estimator method [26].
- Largest Lyapunov Exponent (LLE): LLE indicates the stability of the dynamic system over time. A positive (negative) exponent implies divergence (convergence). The present study computes the average exponential growth of inter-orbit distance (in the phase space) through the prediction error. We estimated LLE using a direct method [27].
- Fractal Dimension (FD): FD is a relative measure of the number of basic building blocks forming the signal. We computed two types of FDs: one from the raw time series (in the time space), and the other from the time-delay reconstructed time series (in the phase space). We used the Petrosian’s algorithm [28] for the former, while we used moments of neighbor distances for the latter.

We determined the optimal time delay at the first local minimum point of auto mutual information. We estimated the

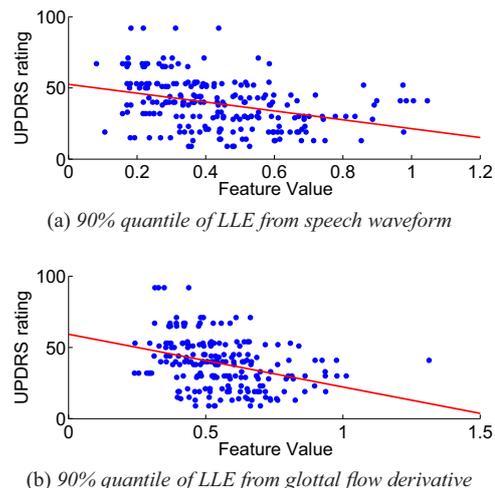


Figure 1: Linear model (red color) overlaid on the scatter plots of the best NTSA feature from speech waveform or glottal flow derivative v.s. the UPDRS label.

minimum embedding dimension, using the Cao’s method [29]. For each time series, we computed CD, LLE, LLE from the smoothed  $p(x)$  plot, where  $p$  is the average exponential growth of the distance of neighboring and  $x$  is the number of time steps. We also computed FDs in both time and phase spaces, resulting in five NTSA features in total. For f0 and RMS-energy time series, we computed the five NTSA features at the utterance level. For the speech waveform, glottal flow and its delta, we first computed the five NTSA features for each of short time window (0.4 sec.) in the speech regions, then computed statistics, such as mean, standard deviation, median, interquartile range, max., min. and [0.1 0.9] quantiles.

Figure 1 illustrates the scatter plots for the individual NTSA features (x-axis) and the UPDRS label (y-axis), only for the task of rapid repetition of syllables as examples. The most correlated features (i.e., 90% quantiles of LLEs) from speech waveform and glottal flow derivative, to the UPDRS label are chosen in Figure 1a and Figure 1b, respectively. The Spearman correlations of the features are  $-0.33$  (speech waveform) and  $-0.34$  (glottal flow derivative). The p-values of the linear models (red color in Figure 1) are  $1.6 \times 10^{-7}$  (speech waveform) and  $1.7 \times 10^{-7}$  (glottal flow derivative). The statistically significant negative relations between UPDRS rating and the features indicate less fluctuation of signals within short-time segments for patients with high PD severity.

## 5. System Development

We tested several schemes such as stacked generalization [30], ranking [31] and regression [32] for predicting the Parkinson’s severity. The best system in our cross-validation experiments was a Support Vector Regressor (SVR), trained on a selected set of features. We pool all the proposed features and perform a 3-stage filter feature selection including: (i) discarding features with Spearman correlation with target labels below a threshold (on the training set), (ii) if a set of features is highly correlated, we retain only one feature in the group, and (iii) removing features with highly dissimilar distributions across training and evaluation (development or test) sets. For the step (ii), we group features with correlations amongst themselves above a threshold (empirically set to 0.98) and retain

	Devel set	
	Spearman	Pearson
*Baseline (C=10 <sup>-3</sup> )	0.49	-
*Baseline (C=10 <sup>-5</sup> )	0.37	-
Fusion	0.51	0.24
Joint rating 1	0.50	0.59
Joint rating 2	<b>0.53</b>	<b>0.63</b>
	Test set	
	Spearman	Pearson
*Baseline (C=10 <sup>-3</sup> )	0.24	-
*Baseline (C=10 <sup>-5</sup> )	0.39	-
Joint rating 2 on Baseline (C=10 <sup>-5</sup> )	<b>0.43</b>	0.37
Joint rating 2 on Fusion	0.42	<b>0.44</b>

Table 2: Spearman and Pearson correlations of the final systems for the PC Sub-Challenge. ‘Joint rating 1’ refers to the within-task within-cluster joint rating. ‘Joint rating 2’ refers to the within-merged-cluster joint rating. Results with \* are adopted from the baseline paper. The cell of the best performance is highlighted.

the feature which is most normally distributed according to the Kolmogorov-Smirnoff test [33]. The step (iii) was designed to take care of mismatch between the training and test sets. Empirical study of the data revealed that several features have different distributions on the training, the development and the test sets. For a feature  $f$ , we compute its mean and standard deviations on train ( $\mu_{\text{train}}^f, \sigma_{\text{train}}^f$ ), development ( $\mu_{\text{dev}}^f, \sigma_{\text{dev}}^f$ ) and testing set ( $\mu_{\text{test}}^f, \sigma_{\text{test}}^f$ ). While training a model to evaluate performance on development set, the feature  $f$  is retained only when  $\mu_{\text{train}}^f - \sigma_{\text{train}}^f < \mu_{\text{dev}}^f < \mu_{\text{train}}^f + \sigma_{\text{train}}^f$ . A similar operation is performed while training a model for evaluation on the test set. Note that this may lead to a different set of features getting selected during evaluation on development and test sets. Therefore, we train two separate SVRs to evaluate performance on development and testing set. Finally, we performed two joint rating approaches using clustering and cluster-merging outputs. One (‘Joint rating 1’ in Table 2) is to impose the median of each cluster in individual speech tasks, to all files of the cluster. We refer to this as the within-task within-cluster joint rating. The other (‘Joint rating 2’ in Table 2) is to impose the median of the optimal task, which has the minimum sum of cross-task distances, to all merged clusters. We refer to this as the within-merged-cluster joint rating.

In the system on the EC Sub-Challenge, the classification was done by training an SVM with polynomial kernel (fifth order) on either the functionals or the i-vectors extracted on the training utterances using the annotated food types as class labels. System combination was performed by linear fusion of the SVM output posteriors. Just as in [9], we applied a leave-one-speaker-out cross-validation (LOSO-CV) to find the optimal parameter settings for the test set.

## 6. Results and discussion

### 6.1. Parkinson’s Condition

Table 2 shows results for final predictions on the development set and preliminary results on the test set. We generated the final prediction labels using early feature fusion scheme and joint rating approaches. The correlations of ‘Joint rating 2’ are higher than those of ‘Fusion’ on the development set, suggesting that

	Train CV	Test set
Baseline	61.3	65.9
Functionals	65.3	-
I-vectors	75.7	-
System fusion	<b>76.2</b>	<b>74.6</b>

Table 3: Spearman correlation of the final fusion systems for the EC Sub-Challenge. The cell of the best performance is highlighted.

within-merged-cluster joint rating approach reduces the prediction errors successfully. Although the Spearman correlation of ‘Joint rating 1’ is slightly lower than that of ‘Fusion,’ Pearson correlation of ‘Joint rating 1’ is significantly higher than that of ‘Fusion,’ suggesting that within-type within-cluster joint rating approach is useful for prediction performance. On the test set, we applied the within-merged-cluster approach on the baseline feature sets and the fusion system. Results indicate that our final system (‘Fusion’ + ‘Joint rating 2’) is capable of generating more accurate rating than the baseline system. Although joint rating on ‘Fusion’ shows slightly lower Spearman correlation than joint rating on the baseline features, the gain for Pearson correlation is significant. These results suggest the usefulness of our proposed features for better severity rating of PD. A higher-Pearson and lower-Spearman case (‘Joint rating 1’ in Table 2) indicates that the predicted labels have better linear relation with the true labels, but the relative ordering of predicted values is worse.

### 6.2. Eating Condition

We used *Unweighted Average Recall* (UAR) as the evaluation metric of our system, defined as the unweighted (by number of utterances in each class) mean of the percentage correctly classified in the diagonal of the confusion matrix. Table 3 presents the numbers on the training set (using a 20-fold LOSO-CV strategy) and test set. From Table 3 we conclude that the i-vector system provides the highest classification accuracy in terms of food types. We obtained an additional increase in performance by fusing the i-vector system with the SVM posteriors from the functionals (bottom row in Table 3).

## 7. Summary and future works

This paper addresses the PC Sub-Challenge mainly and proposes novel features for capturing various paralinguistic cues, in particular irregular and atypical aspects in speech signal. We use the features to train models which do not only account for a mismatch in the data splits, but also incorporate acoustic similarity of utterances into the final severity rating. Our results show that the proposed features and models improve the baseline system. The potential of the proposed features is further supported by their performance in the EC Sub-Challenge.

Our future works include further investigation into the nature of the two Sub-Challenge problems. For instance, the Parkinson’s severity rating does not only present a challenge of reliable prediction, but also accounts for a mismatch in the data splits, e.g., different recording conditions and patients. Addressing these issues using machine learning techniques, e.g., features adaptation [34] and semi-supervised learning methods [35], is a part of our future works.

## 8. Acknowledgements

Work supported by NSF, NIH and DoD.

## 9. References

- [1] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of parkinson's disease progression by non-invasive speech tests," *Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 884–893, 2010.
- [2] A. Bayestehtashk, M. Asgari, I. Shafran, and J. McNames, "Fully automated assessment of the severity of parkinson's disease from speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 172–185, 2015.
- [3] A. Sepulveda, G. Castellanos-Dominguez, and R. C. Guido, "Time-frequency relevant features for critical articulators movement inference," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 2802–2806.
- [4] S. Skodda, A. Flasskamp, and U. Schlegel, "Instability of syllable repetition as a model for impaired motor processing: is parkinsons disease a rhythm disorder?" *Journal of Neural Transmission*, vol. 117, no. 5, pp. 605–612, May 2010. [Online]. Available: <http://link.springer.com/10.1007/s00702-010-0390-y>
- [5] L. O. Ramig and C. Dromey, "Aerodynamic mechanisms underlying treatment-related changes in vocal intensity in patients with parkinson disease," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 4, pp. 798–807, 1996.
- [6] G. Vaziri, F. Almasganj, and R. Behroozmand, "Pathological assessment of patients' speech signals using nonlinear dynamical analysis," *Computers in Biology and Medicine*, vol. 40, no. 1, pp. 54–63, 2010.
- [7] T. Bocklet, E. Noth, G. Stemmer, H. Ruzickova, and J. Ruzs, "Detection of persons with parkinson's disease by acoustic, vocal, and prosodic analysis," in *Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, Dec 2011, pp. 478–483.
- [8] J. Ruzs, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinsons disease," *The Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 350–367, 2011.
- [9] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönl, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinsons & Eating Condition," in *Proceedings of Interspeech*. ISCA, 2015.
- [10] J. R. Orozco-Arroyave, J. D. Arias-Londono, J. F. Vargas-Bonilla, M. C. Gonzalez-Rtiva, and E. Nöth, "New Spanish speech corpus database for the analysis of people suffering from Parkinsons disease." [Online]. Available: [http://www.lrec-conf.org/proceedings/lrec2014/pdf/7\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2014/pdf/7_Paper.pdf)
- [11] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE - The Munich versatile and fast open-source audio feature extractor," in *ACM Multimedia*, Firenze, Italy, 2010, pp. 1459–1462.
- [12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec 2011, iEEE Catalog No.: CFP11SRW-USB.
- [13] "1997 Spanish Broadcast News Speech (HUB4-NE) LDC98S74," 1998, web Download.
- [14] W. Wang, P. Lu, and Y. Yan, "An improved hierarchical speaker clustering," *ACTA Acustica*, vol. 33, pp. 9–14, 2008.
- [15] K. Han, S. Kim, and S. Narayanan, "Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization," *Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1590–1601, November 2008.
- [16] J. Kim, N. Kumar, A. Tsiartas, and S. Narayanan, "Intelligibility classification of pathological speech using fusion of multiple high level descriptors," in *Proceedings of Interspeech*. ISCA, 2012, pp. 534 – 537.
- [17] M. V. Segbroeck, R. Travadi, C. Vaz, J. Kim, M. P. Black, A. Potamianos, and S. S. Narayanan, "Classification of cognitive load from speech using an i-vector framework," in *Proceedings of Interspeech*. ISCA, 2014, pp. 751 – 755.
- [18] Y. Shao, Z. Jin, D. L. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *Proceedings of ICASSP*, 2009.
- [19] M. Kleinschmidt, "Spectro-temporal gabor features as a front end for ASR," in *Proc. Forum Acusticum Sevilla*, 2002.
- [20] M. Van Segbroeck, A. Tsiartas, and S. Narayanan, "A robust front-end for VAD: Exploiting contextual, discriminative and spectral cues of human voice," in *Proceedings of Interspeech*. ISCA, 2013.
- [21] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.05) [computer program]," 2009, retrieved Jan 11, 2011. [Online]. Available: <http://www.praat.org/>
- [22] K. Audhkhasi, A. M. Zavou, P. G. Georgiou, and S. S. Narayanan, "Theoretical analysis of diversity in an ensemble of automatic speech recognition systems," *Transactions on Audio, Speech & Language Processing*, vol. 22, no. 3, pp. 711–726, 2014.
- [23] M. Van Segbroeck, R. Travadi, and S. S. Narayanan, "UBM fused total variability modeling for language identification," in *Proceedings of Interspeech*. ISCA, 2014.
- [24] M. Van Segbroeck, R. Travadi, C. Vaz, J. Kim, M. P. Black, A. Potamianos, and S. S. Narayanan, "Classification of cognitive load from speech using an i-vector framework," in *Proceedings of Interspeech*. ISCA, 2014.
- [25] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, no. 2-3, pp. 109 – 118, 1992.
- [26] T. Schreiber, "Interdisciplinary application of nonlinear time series methods," *Physics reports*, vol. 308, no. 1, pp. 1–64, 1999.
- [27] U. Parlitz, "Nonlinear time series analysis," in *Proceedings of the Third International Specialist Workshop on Nonlinear Dynamics of Electronic Systems*. University College Dublin, Ireland: NDES'95, Jul 1995, pp. 179 – 192.
- [28] A. Petrosian, "Kolmogorov complexity of finite sequences and recognition of different preictal EEG patterns," in *Proceedings of the Eighth IEEE Symposium on Computer-Based Medical Systems*, Jun 1995, pp. 212–217.
- [29] L. Cao, "Practical method for determining the minimum embedding dimension of a scalar time series," *Physica D: Nonlinear Phenomena*, vol. 110, no. 1, pp. 43–50, 1997.
- [30] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.
- [31] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, pp. 80–83, 1945.
- [32] T. P. Ryan, *Modern Regression Methods*, 2nd ed. New York, NY, USA: Wiley-Interscience, 2008.
- [33] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.
- [34] S. Tibrewala and H. Hermansky, "Multi-band and adaptation approaches to robust speech recognition," in *EUROSPEECH*, 1997.
- [35] O. Chapelle, B. Schlkopf, and A. Zien, *Semi-Supervised Learning*, 1st ed. The MIT Press, 2010.