# Parkinson's Condition Estimation using Speech Acoustic and Inversely Mapped Articulatory Data

*Seongjun Hahm*[1], *Jun Wang*[1,2,3]

[1]Speech Disorders & Technology Lab, Department of Bioengineering
[2]Callier Center for Communication Disorders
University of Texas at Dallas, Richardson, Texas, United States
[3]University of Texas Southwestern Medical Center, Dallas, Texas, United States
{seongjun.hahm, wangjun}@utdallas.edu

## Abstract

Parkinson's disease is a neurological disorder that affects patient's motor function including speech articulation. There is no cure for Parkinson's disease. Speech and motor function declines as the disease progresses. Automatic assessment of the disease condition may advance the treatment of Parkinson's disease with objective, inexpensive measures. Speech acoustics, which can be easily obtained from patients, has been used for automatic assessment. The use of information in motor function of articulator (e.g., jaw, tongue, or lips) has rarely been investigated. In this paper, we proposed an approach of automatic assessment of Parkinson's condition using both acoustic data and acoustically-inverted articulatory data. The quasi-articulatory features were obtained from the Parkinson's acoustic speech data using acoustic-to-articulatory inverse mapping. Support vector regression (SVR) and deep neural network (DNN) regression were used in the experiment. Results indicated adding articulatory data to acoustic data can improve the performance of using acoustic data only, for both SVR and DNN. In addition, deep neural network outperformed support vector regression on the same data features measured with Pearson correlation but not with Spearman correlation. The implications of our approach with further improvement were discussed.

**Index Terms**: Parkinson's disease, acoustic-to-articulatory inverse mapping, support vector regression, deep neural network

## 1. Introduction

Parkinson's disease (PD) is one of the most common neurodegenerative disorders that affects one's motor function, and therefore impairs the speech. Parkinson's disease affects the life of about one million in the United States and about five millions worldwide [1]. Parkinson's disease is a result of the loss of dopamine-producing brain cells [2]. The causes of Parkinson's disease are still unknown currently and there is no cure [1, 3]. Patient's and their family's lives are severely impacted due to the disease. Treatment of Parkinson's disease is a huge economic burden for taxpayers and patients [4].

Current assessment techniques for Parkinson's disease are relying on human experts and thus expensive, subjective, and time-consuming [4, 5]. Unified Parkinson's Disease Rating Scaling (UPDRS) is the most widely used measure for evaluating the symptom severity [6, 7], which consists of five sections of evaluation including motor, mentation, mood, behavioral, self-evaluation [8]. The rating score range is from 0 to 176, where 0 represents completely healthy and 176 is totally disabled [9]. Objective and automatic measures for PD symptom severity assessment is highly needed.

Recent advances have been made for automatic and objective assessment of PD severity using speech signals [9, 10, 11, 12]. Speech signals can be collected conveniently in a clinical environment or even at home through telephone or cell phone [13]. Acoustic speech signals have also been used for change-point detection in PD [14]. Common voice symptoms due to PD include reduced loudness, monotone, hoarseness, breathiness (noise), and vocal tremor [15]. The basic idea of these approaches is to extract features that represent those symptoms from acoustic data and then evaluate the severity.

Articulatory motor function decline due to PD, however, has rarely been used for automatic symptom severity estimation. There are four primary motor symptoms of PD including 1) tremor, or trembling in hands, arms, legs, jaw, and face; 2) rigidity, or stiffness of the limbs and trunk; 3) bradykinesia, or slowness of movement; and 4) postural instability, or impaired balance and coordination [2]. Trembling in jaw and slowness of articulators lead to imprecise articulation, the direct causes of the impaired speech. These symptoms in articulation motivated a novel approach for automatic PD severity estimation from articulatory data. For example, the movement patterns of jaw, tongue, and lips can be used together with acoustic speech data for automatic Parkinson's condition estimation.

Despite the logistical difficulty for articulatory data collection [16], articulatory data could be inversely mapped from acoustic data [17]. For example, Electromagnetic Articulograph (EMA) is one of the currently used techniques for collecting tongue and lip movement data during speech [18]. EMA records articulatory movement data by attaching small wired sensors on the surface of jaw, tongue, and lips [19]. Articulatory and associated acoustic data that have been collected using EMA could be used to build an inverse mapping model. This model can then be used for deriving articulatory data from acoustic data collected from PD patients. Acoustic-to-articulatory inverse mapping has been proven feasible with small errors in recent studies [20, 21] and the mapping can be speaker-independent [22].

As a participation in the Interspeech 2015 Computation Paralinguistics Challenge (Parkinson's Condition sub-challenge) [23], this paper investigated the use of inversely mapped articulatory data in Parkinson's Condition estimation. To our best knowledge, this is the first attempt of Parkinson's Condition (PC) estimation using articulatory data. A publicly available articulatory and acoustic data, MOCHA-TIMIT [24], was used

to build the inverse mapping model. Then the provided acoustic data set was inversely mapped to articulatory data. Articulatory features were extracted from the articulatory data and were used together with acoustic data to test the PC estimation performance. Our results will be compared with the baseline results using support vector regression (also provided in [23]). Moreover, deep neural network (DNN) has recently attracted attention of researchers because it (together with hidden Markov model, HMM) outperformed the long-standing approach Gaussian mixture model-HMM in speech recognition [25] and other domains [26]. Deep neural network was also used in this experiment as compared to the performance obtained with support vector regression (SVR).

# 2. Method

The core procedure of using acoustic-to-articulatory mapping for PC estimation is in three steps: (1) to train an acoustic-to-articulatory inverse-mapping model using an available data set, (2) then to apply the model to inversely map Parkinson's speech (acoustic) data to articulatory data, (3) finally use the combined acoustic and quasi-articulatory features for PC estimation. We hypothesized that PC estimation performance would be improved with the additional quasi-articulatory features, compared with that using acoustic features only.

## 2.1. Acoustic-to-Articulatory Inverse Mapping

### *2.1.1. Dataset for Inverse Mapping*

MOCHA(Multi-CHannel Articulatory)-TIMIT, a publicly available database with synchronously recorded acoustic and articulatory data, was used to train the inverse mapping model. MOCHA-TIMIT data set consists of simultaneous recordings of speech, articulatory data from 2 British English speakers (1 male - MSAK0 and 1 female - FSEW0) [24]. There are in total 920 sentences (extracted from TIMIT database).

The articulatory data was collected using an Electromagnetic Articulograph (EMA, Carstens Medizinelektronik GmbH, Germany) by attaching sensors to upper lip (UL), lower lip (LL), upper incisor (UI), lower incisor (LI or Jaw), tongue tip (TT), tongue blade (TB), tongue dorsum (TD), and velum (V) with 500 Hz sampling rate (downsampled to 100 Hz). Each sensor has two dimensions, *x* (front-back) and *y* (vertical) trajectories, since lateral (left-right) movements are not significant in speech production of healthy speakers [16]. The silences before and after the utterance were removed. According to our experience, UI does not involve significant movement and therefore was not used in this experiment. Thus, the acoustic data and the 14-dimensional articulatory motion data obtained from LI, V, UL, LL, TT, TB, and TD were used in this experiment.

Table 1 lists the sensors (flesh points on articulators) that were used in both the inverse-mapping and the PC estimation.

### *2.1.2. Inverse Mapping*

Deep neural network (DNN) regression was used as the speaker-independent inverse mapping model [21]. The DNN composed of 3-hidden layers which has 256 nodes at each layer. The input of DNN is 225 concatenated MFCC feature vectors: 3 consecutive (previous, current, and succeeding frames) 75-dimensional MFCC feature vectors (25 MFCC + $\Delta$ + $\Delta\Delta$). The output of DNN is the estimated 14-dimensional articulatory (EMA) feature vectors. For training and regression, we used KALDI speech recognition toolkit [27]. A low pass filter (20 Hz cutoff frequency) was used to smooth the data after the inverse mapping [21].

Table 1: *Flesh points on articulators*

| Sensor | Full Name |
|---|---|
| LI | Lower Incisor (Jaw) |
| UL | Upper Lip |
| LL | Lower Lip |
| TT | Tongue Tip |
| TB | Tongue Blade |
| TD | Tongue Dorsum |
| V | Velum |

The results of inverse mapping was evaluated using the root-mean-square-error (RMSE) between the measured (original) and the inversely mapped articulation motion paths of all sensors. A better (accurate) performance of inverse mapping benefits the Parkinson's Condition estimation performance.

## 2.2. Quasi-Articulatory Features

The trained inverse mapping model was applied on the Parkinson's speech data to generate articulatory data. Quasi-articulatory features were then extracted from the inversely mapped Parkinson's articulatory data. The quasi-articulatory features were then used together with acoustic features for PC estimation.

The script provided in [23] was modified (70 ms window size and 35 ms frame shift) and used to extract quasi-articulatory features from inversely mapped articulatory motion data. The script automatically extracts up to 6,373 pre-defined acoustic features, including jitter, shimmer, and MFCC. However, low frequency articulatory data do not contain these information. Thus, we disabled the features below when using the tool to extract quasi-articulatory features:

*Jitter, Shimmer, logHNR, Rfilt, Rasta, MFCC, Harmonicity, and Spectral Rolloff.*

Thus, for each dimension (*y* or *z*) of a sensor, 1,200 features were extracted. In total, 23,173 features (6,373 acoustic feature + 2,400 articulatory features $\times$ 7 sensors) were used to test our Parkinson's Condition estimation approaches. In addition, 2,400 articulatory features for selected sensors were added to the 6,373 acoustics features for Parkinson's Condition estimation. This additional test will help to understand the contribution of quasi-articulatory features from individual sensors.

## 2.3. Support Vector Regression

Support vector regression, a regression based on support vector machine [28], was used as the baseline approach. SVR is a soft-margin regression technique that depends only on a subset of the training data, because the cost function for building the model does not care about training points that are beyond the margin [29], which is the similar of SVM. Details on the introduction of SVR can be found in [30].

We reproduced the results that are provided in [23] as a practice for understanding the dataset. Results obtained using our proposed approaches will be compared with these obtained using SVR.

## 2.4. Deep Neural Network based Regression

Recently, DNN-HMM showed the significant performance improvement compared with the long-standing approach Gaussian

mixture model (GMM)-HMM [25, 31, 32, 33] in speech recognition and other applications [26, 34]. In this paper, DNN training approach based on restricted Boltzmann machines (RBMs) [35] was used for regression.

The DNN (stacked RBMs) is subsequently fine-tuned using backpropagation algorithm. A detailed explanation and further discussion of the DNN can be found in [31, 32, 35].

The structure of the DNN (3 hidden layers with 512 nodes in each layer) used for Parkinson's Condition estimation is similar as that used for acoustic-to-articulatory mapping (Section 2.1), except the input and output layers. The input features in the Parkinson's Condition estimation is all extracted features (acoustic + articulatory features). The output layer had only one node, the estimated PC score.

## 3. Experimental Design

As stated previously, the goal of this project is to test if adding quasi-articulatory features from the inversely mapped articulatory data can improve the Parkinson's Condition estimation performance. Thus all extracted quasi-articulatory features were added to the acoustic feature set to test if there is a performance improvement over that using acoustic features only.

Moreover, our inverse mapping model generated articulatory data in a form of 7 sensors (Table 1). Quasi-articulatory features extracted from selected individual sensor's data were added to the acoustic features to identify which sensor(s) may perform better than others. We hypothesize LI (Jaw) might obtain the best performance, because trembling of jaw is one of the symptoms in Parkinson's disease [9].

Finally, DNN regression was tested in the Parkinson's Condition estimation and the performance was compared with that obtained using SVR.

### 3.1. Parkinson's Data Set

The data set provided in [3] was used in the experiments. The data set consists speech data collected from a total 50 patients with Parkinson's disease (25 males and 25 females). Each participant performed a total of 42 speech tasks including 24 isolated words, 10 phrases, one reading task, one monologue, and the rapid repetition of /pa-ta-ka/, /pa-ka-ta/, and /pe-ta-ka/ [23]. A set of up to 6,373 features have been extracted from and are provided with the raw audio files. Details of the data set, including participants' information and UPDRS scores, are provided in [3, 23].

### 3.2. Outcome Measures

Two correlations, Spearman and Pearson, were used to evaluate the performance of our approaches, although only Spearman correlation was provided in [23]. Pearson correlation is more sensitive than Spearman correlation for outliers. Using both correlations may provide more detailed information for interpreting our experimental results [36]. A higher correlation between the estimated condition and the actual condition (UPDRS) indicates a better performance.

## 4. Results and Discussion

### 4.1. Acoustic-to-articulatory Inverse Mapping

Table 2 gives the performance of acoustic-to-articulatory inverse mapping using DNN with or without low pass filtering (LPF). The root-mean-square-error (RMSE) values are an average of the RMSEs in 5-fold cross validation, for each dimension of each sensor.

Table 2: *Inverse-mapping results on MOCHA-TIMIT data set: Overall Root-Mean-Squared Errors (RMSE; mm) of measured (original) and estimated motion paths.*

| Sensors | w/o LPF | w/ LPF |
|---|---|---|
| LI_x | 0.72 | 0.68 |
| UL_x | 0.75 | 0.72 |
| LL_x | 1.15 | 1.08 |
| TT_x | 2.34 | 2.08 |
| TB_x | 2.16 | 1.92 |
| TD_x | 2.02 | 1.82 |
| V_x | 0.48 | 0.45 |
| LI_y | 1.10 | 1.02 |
| UL_y | 1.12 | 1.04 |
| LL_y | 2.13 | 1.97 |
| TT_y | 2.53 | 2.28 |
| TB_y | 2.04 | 1.82 |
| TD_y | 2.15 | 1.90 |
| V_y | 0.76 | 0.71 |
| Average | 1.53 | 1.39 |

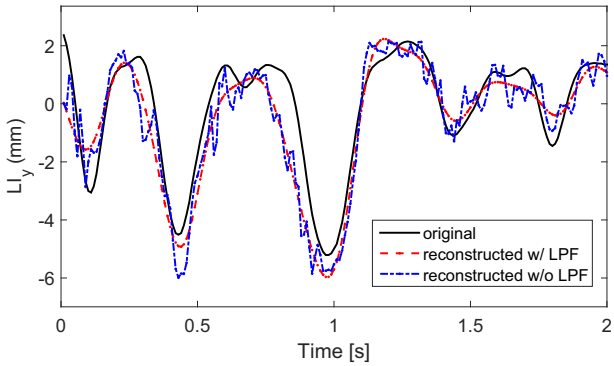


Figure 1: *Example of reconstructed articulatory motion paths of LI (Jaw) with and without low-pass filtering (LPF).*

Different RMSEs were obtained from different sensors. LI (Jaw) and velum (V) obtained smaller RMSEs than other sensors. Tongue, including TT, TB, and TD obtained largest RMSEs. This finding may indicate that jaw and velum contains less variation and tongue movements have more variation in speech production.

Figure 1 illustrates three articulatory motion paths of Jaw producing a sentence. The black solid curve is the original; the blue dashed is the inversely mapped motion path before applying LPF; the red dashed is inversely mapped motion path after LPF was applied. Although the three curves are not perfectly overlapped, the inversely mapped articulatory paths generally follow the original path, which visually verified the effectiveness of the inverse mapping model.

### 4.2. Parkinson's Condition Estimation

Although inverse-mapping with LPF outperformed that without LPF (Table 2), our preliminary tests indicated that using LPF causes slightly worse performance in Parkinson's Condition estimation. It is possibly because LPF removed some useful information that impacts the quasi-articulatory feature extraction. Thus, LPF was not used when converting the Parkin-

Table 3: *Parkinson's Condition estimation using SVR and DNN with acoustic and articulatory features from individual sensors.*

| Method | Feature Set | Correlation Coefficient | |
|---|---|---|---|
| | | Pearson | Spearman |
| SVR | Acoustic (Baseline) | 0.3457 | 0.4916 |
| | Acoustic + UL | 0.3730 | 0.4924 |
| | Acoustic + LL | 0.4112 | 0.4941 |
| | Acoustic + LI (Jaw) | 0.3913 | **0.5027** |
| | Acoustic + ALL | **0.4457** | 0.4625 |
| DNN | Acoustic | 0.4721 | 0.4632 |
| | Acoustic + UL | 0.4714 | 0.4580 |
| | Acoustic + LL | 0.4724 | 0.4567 |
| | Acoustic + LI (Jaw) | 0.5010 | **0.4895** |
| | Acoustic + ALL | **0.5113** | 0.4854 |

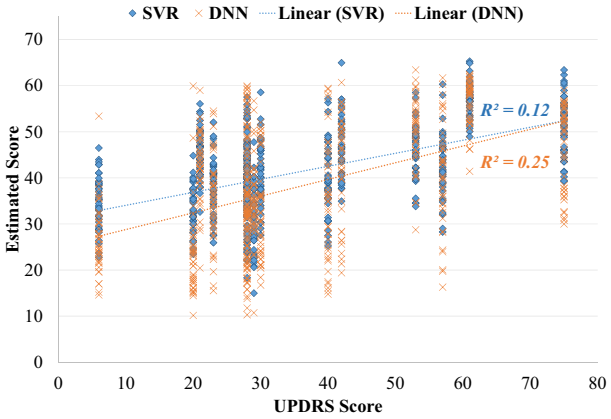


Figure 2: *Scatter plot of actual UPDRS score and estimated scores using SVR and DNN (better viewed in color).*

son's speech (acoustic) data to articulatory data.

Table 3 gives the Parkinson's Condition estimation performance (Spearman and Pearson correlations) in a combination of features and classifiers. Feature combinations include (1) acoustic only (baseline), (2) acoustic + UL, (3) acoustic + LL, (4) acoustic + LI (Jaw), and (5) acoustic + ALL features (7 sensors) using the development data set [23]. All Pearson and Spearman correlations in Table 3 using SVR were obtained with complex parameter $C = 10^{-3}$, the best out of other $C$ values.

One important finding based on Table 3 is adding quasi-articulatory features improved the performance (as measured either with Pearson or Spearman correlations) than that using acoustic features only, for both SVR and DNN regression.

In addition, Table 3 indicates adding even articulatory features from a single visible sensor (e.g., UL, LL, or Jaw) can improve Parkinson's condition estimation performance. LI (Jaw) obtained the best performance among other single sensors.

This finding may have great potential for the development of a portable visual-speech Parkinson's Condition estimator, since jaw movement can be easily tracked using a webcam. For example, a smart phone can be used to record speech sounds and track the jaw movement to evaluate the Parkinson's condition.

When comparing the performances of SVR and DNN regression, different results were obtained with Pearson and Spearman correlations (Table 3). DNN regression outperformed SVR when using Pearson correlation, while SVR outperformed

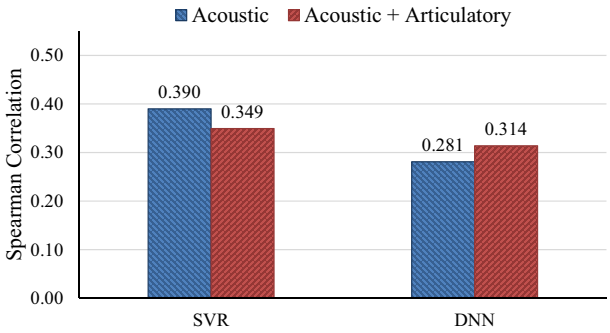


Figure 3: *Parkinson's Condition estimation performance using different features on Test dataset (for the PC Challenge). The performance was measured using Spearman correlation.*

DNN regression when using Spearman correlation. This is mainly because Spearman correlation is less sensitive for outliers [23], thus gave higher values for SVR. In this experiment using the development dataset, SVR produced at least two outliers, while DNN produced no outliers.

Figure 2 shows the scatter plot of the actual UPDRS score and the estimated scores using SVR and DNN regression (outliers not displayed). A linear regression suggested DNN regression ($R^2 = 0.25$) outperformed SVR ($R^2 = 0.12$).

Figure 3 provides the results using the Test set, which is for the competition in the Interspeech 2015 PC Challenge. Our proposed approach (DNN + acoustic features + quasi-articulatory features) obtain results that are worse than the baseline results provided in [23]. However, based on previous discussion, we still think our proposed approach is better than the baseline approach (SVR with acoustic features) [23].

Although the inverse mapping model trained using healthy data was used, encouraging results were obtained. We believe that the PC estimation performance can be further improved when the model is built from real Parkinson's data. This finding motivated an articulatory data collection from patients with Parkinson's disease.

Our approaches using quasi-articulatory features in this experiment are data-driven (rely on SVR or DNN regression), although we mentioned trembling of jaw is one of the primary symptom in Parkinson's. Interpretable model (e.g., with a derived feature for jaw trembling) may further improve the performance and advance the understanding of Parkinson's symptom.

## 5. Conclusions and Future Work

In this paper, we proposed a novel approach for automatic Parkinson's condition estimation (using acoustic speech data and inverted articulatory data). The approach was tested using the data sets provided in [23]. Experimental results indicated the performance improvement of adding quasi-articulatory features, particularly from jaw.

Future directions include (1) Parkinson's condition estimation using the quasi-articulatory feature only, (2) collecting real articulatory data from patients with Parkinson's disease for severity estimation, and (3) tuning the structure of DNN to improve the performance of inverse-mapping and Parkinson's Condition estimation.

## 6. Acknowledgment

We would like to thank the organizers of the Interspeech 2015 Computational Paralinguistics Challenge (PC Sub-challenge).

# 7. References


[1] L. M. de Lau and M. M. Breteler, "Epidemiology of Parkinson's disease," *The Lancet Neurology*, vol. 5, pp. 525–535, 2006.

[2] National Institute of Neurological Disorders and Stroke. (2015, March) MNINDS Parkinson's Disease Information Page. [Online]. Available: "http://www.ninds.nih.gov/disorders/parkinsons_disease/parkinsons_disease.htm"

[3] J. Orozco-Arroyave, J. Arias-Londono, J. Vargas-Bonilla, M. Gonzlez-Rtiva, and E. Nth, "New spanish speech corpus database for the analysis of people suffering from parkinsons disease," in *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, 2014, p. 342347.

[4] S. L. Kowal, T. M. Dall, R. Chakrabarti, M. V. Storm, and A. Jain, "The current and projected economic burden of Parkinson's disease in the United States," *Movement Disorders*, vol. 23, pp. 311–318, 2013.

[5] C. A. Haaxma, B. R. Bloem, G. F. Borm, and M. W. Horstink, "Comparison of a timed motor test battery to the unified Parkinson's disease rating scale-iii in Parkinson's disease," *Movement Disorders*, vol. 23, pp. 1707–1717, 2008.

[6] G. T. Stebbins and C. G. Goetz, "Factor structure of the Unified Parkinson's Disease Rating Scale: Motor Examination section," *Movement Disorders*, vol. 13, pp. 633–636, 1998.

[7] C. Ramaker, J. Marinus, A. M. Stiggelbout, and V. H. B.J., "Systematic evaluation of rating scales for impairment and disability in Parkinson's disease," *Movement Disorders*, vol. 17, pp. 867–876, 2002.

[8] C. G. Goetz et al., "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Process, format, and clinimetric testing plan," *Movement Disorders*, vol. 22, pp. 41–47, 2007.

[9] T. A., L. M.A., M. P.E., and R. L.O., "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*, vol. 8, pp. 842–855, 2011.

[10] M. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinsons disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, pp. 1015–1022, 2009.

[11] J. V. squez Correa, J. Orozco-Arroyave, J. Arias-Londono, J. F. Vargas-Bonilla, and E. N. th, "New computer aided device for real time analysis of speech of people with Parkinsons disease," *Fac. Ing. Univ. Antioquia*, vol. 72, pp. 87–103, 2014.

[12] A. Tsanas, M. Little, P. McSharry, J. Spielman, and L. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinsons disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, pp. 1264–1271, 2012.

[13] C. G. Goetz et al., "Testing objective measures of motor impairment in early Parkinsons disease: feasibility study of an at-home testing device," *Movement Disorders*, vol. 24, pp. 551–556, 2007.

[14] R. Cmejla, J. Rusz, P. Bergl, and J. Vokral, "Bayesian changepoint detection for the automatic assessment of fluency and articulatory disorders," *Speech Communication*, vol. 558, pp. 178–189, 2013.

[15] R. Pahwa and K. L. (Eds.), *Handbook of Parkinsons Disease, 4th Edition*. Informa Healthcare, 2007.

[16] J. Wang, J. Green, A. Samal, and Y. Yunusova, "Articulatory distinctiveness of vowels and consonants: A data-driven approach," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 5, pp. 1539–1551, 2013.

[17] A. B. Youssef, P. Badin, G. Bailly, and P. Heracleous, "Acoustic-to-articulatory inversion using speech recognition and trajectory formation based on phoneme hidden markov models," in *Proc. of INTERSPEECH*, 2009, pp. 2255–2258.

[18] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[19] J. Green, J. Wang, and D. L. Wilson, "Smash: A tool for articulatory data processing and analysis," in *Proc. of INTERSPEECH*, Vancouver, Canada, 2013, pp. 1331–1335.

[20] T. Toda, A. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with gaussian mixture model," in *Proc. of INTERSPEECH*, 2004, pp. 1129–1132.

[21] C. Canevari, L. Badino, L. Fadiga, and G. Metta, "Relevance-weighted-reconstruction of articulatory features in deep-neural-network-based acoustic-to-articulatory mapping," in *Proc. of INTERSPEECH*, Lyon, France, 2013, pp. 1297–1301.

[22] P. K. Ghosh and S. S. Narayanan, "A subject-independent acoustic-to-articulatory inversion," in *Proc. of ICASSP*, 2004, pp. 4624–4627.

[23] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. H. nig, J. R. Orozco-Arroyave, E. N. th3, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinsons & Eating Condition," in *Proc. of INTERSPEECH*, 2015, p. In press.

[24] A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *Proc. of ICSLP*, Beijing China, 2000, pp. 145–148.

[25] C. Canevari, L. Badino, L. Fadiga, and G. Metta, "Cross-corpus and cross-linguistic evaluation of a speaker-dependent DNN-HMM ASR system using EMA data," in *Proc. of Workshop on Speech Production in Automatic Speech Recognition*, Lyon, France, 2013.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, and V. K., "The Kaldi speech recognition toolkit," in *Proc. of ASRU*, Waikoloa, USA, 2011, pp. 1–4.

[28] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[29] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. N. Vapnik, *Support Vector Regression Machines*. MIT Press, 1997, vol. 9.

[30] A. J. Smola and B. Schlkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[31] A.-R. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[32] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[33] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams *et al.*, "Recent advances in deep learning for speech research at Microsoft," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 8604–8608.

[34] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 2553–2561.

[35] G. Hinton, "A practical guide to training restricted Boltzmann machines," *Momentum*, vol. 9, no. 1, p. 926, 2010.

[36] J. L. Myers, A. Well, and R. F. Lorch, *Research design and statistical analysis*. Routledge, 2010.