INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Interpretable Objective Assessment of Dysarthric Speech based on Deep Neural Networks

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

{mingtu, visar, jmliss}@asu.edu

Abstract

Improved performance in speech applications using deep neural networks (DNNs) has come at the expense of reduced model interpretability. For consumer applications this is not a problem; however, for health applications, clinicians must be able to interpret why a predictive model made the decision that it did. In this paper, we propose an interpretable model for objective assessment of dysarthric speech for speech therapy applications based on DNNs. Our model aims to predict a general impression of the severity of the speech disorder; however, instead of directly generating a severity prediction from a high-dimensional input acoustic feature space, we add an intermediate interpretable layer that acts as a bottle-neck feature extractor and constrains the solution space of the DNNs. During inference, the model provides an estimate of severity at the output of the network and a set of explanatory features from the intermediate layer of the network that explain the final decision. We evaluate the performance of the model on a dysarthric speech dataset and show that the proposed model provides an interpretable output that is highly correlated with the subjective evaluation of Speech-Language Pathologists (SLPs).

Index Terms: dysarthric speech, objective assessment, model interpretability, deep neural networks

1. Introduction

There has been a recent trend of applying machine learning techniques, especially those based on deep neural networks (DNNs), to different applications. In most cases, the goal is improving performance; however, in some applications, it is also important to understand why an algorithm made the decision that it did. This is critical in medical applications where clinicians must know whether they can trust the prediction of a machine learning model if they are to make decisions based upon it [1][2]. To answer this question the model should provide the end-user with a final prediction and an "explanation" rather than simply operating as a black-box.

Interpretability is context-specific. We focus on an application area of recent interest to the speech community: objective assessment of dysarthric speech [3]. Objective assessment of dysarthric speech complements subjective assessment in speech therapy. It can be used to detect early signs of neurological disorder[4], and track progress resulting from behavioral or pharmacological intervention. The principal goal of objective assessment is to estimate the perceived severity or intelligibility of dysarthric speech. This is usually done by a data-driven model trained on collected speech samples and labels from clinical experts. Most of current objective assessment systems either focus on more sensitive acoustic features that can better represent the underlying pathology or more advanced machine learning models that can learn a better mapping from the acoustic features to the label [5][6][7][8][9]. However, most current systems are not amenable to interpretation because of the high-dimensional and non-intuitive input feature space. For example, the openSMILE toolbox [10] provides extraction of several thousand features and was used as the baseline system for the Interspeech 2015 Parkinson's disease challenge [3]. However, most of the input features are difficult to explain in terms that a clinician may understand (e.g. by relating them to the typical speech symptoms of Parkinson's disease).

There has been recent interest in interpretable and explainable models in the artificial intelligence community. For example, in [1] the authors propose a locally approximated linear model to interpret any classifiers used for image recognition and text classification. In [11], a novel loss function is proposed to jointly predict the label of an image and give a sentence describing the prediction. In [12] the authors propose a recommender system that simultaneously trains a latent factor model for rating prediction and a topic model for product reviews. The reviews can be used to justify the predicted rating. It is not a coincidence that most of these previous studies focus on computer vision and natural language processing applications since the input feature spaces for both domains are interpretable to most end-users (pixels and words). Speech is an especially difficult case since the input features that seem to work well in different applications (e.g. Mel-Frequency Cepstral Coefficients (MFCC)) are only understood by experts in the field. Our previous work investigates the relationship between ASR performance and interpretable perceptual disturbances of dysarthric speech and serves as a starting point for the way we define interpretability in this paper [13].

In this paper, we propose a new method for interpretable objective assessment of dysarthric speech. Our model is based on DNNs and aims to predict a general impression of the severity of dysarthric speech. Instead of directly mapping the input acoustic features to a target of interest (e.g. severity), we propose to insert an intermediate layer in the DNNs. This intermediate layer is composed of nodes that can be used to explain the final severity decision. The motivation behind this model is the seminal work of Darley, Aronson and Brown (hereafter, DAB) who proposed a prescriptive methodology for the diagnosis of dysarthric speech [14]. In their paper, they identified 38 perceptual dimensions of speech that clinicians should subjectively assess when making a final decision regarding diagnosis. We summarize these dimensions and force our model to learn the final target (severity) *and* an intermediate DAB representation interpretable to most clinicians that work with pathological speech. In addition to providing interpretation of the

http://dx.doi.org/10.21437/Interspeech.2017-1222

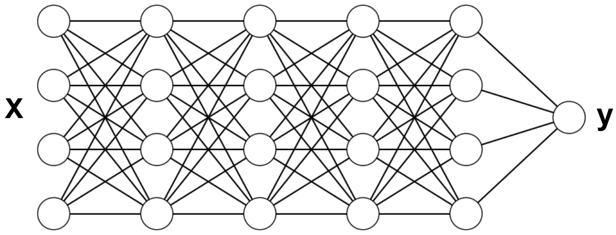

Figure 1: *A notional DNNs for objective assessment of dysarthric speech.*

final output, the intermediate layer also acts as a bottle-neck feature extractor and constrains the solution space of the DNNs by training the DNNs with supervision at both the output layer and the DAB layer. During inference, the model provides an estimate of severity at the output of the network and a set of explanatory features from the DAB layer that aim to justify the final severity decision. The idea of adding an intermediate layer to DNNs has been applied to both ASR and image recognition [15][16][17][18]. However, in these previous studies the added intermediate layer was not interpretable but was used to regularize the network. Our idea of using the DAB layer is also related to hierarchical classifiers for image recognition [19], where the goal is to achieve a trade-off between classification accuracy and specificity.

In the remainder of this paper, we define interpretability for our application, describe the proposed network and two training schemes, and evaluate the performance of the model on a dysarthric speech dataset. The results show that the proposed model provides an interpretable output that is highly correlated with the subjective evaluation of SLPs.

2. Defining Interpretability

When evaluating dysarthric speech, clinicians use the DAB criteria. The 38 perceptual dimensions defined in [14] can be broadly categorized into four groups: perceptual symptoms of atypical nasality, vocal quality, articulatory precision and prosody. Clinicians evaluate speech along some subset of the DAB perceptual dimensions on a 7-point scale (typical to severely atypical). The DAB dimensions are clearly understood by both speech-language pathologists and neurologists and we posit that they serve as a good way to define interpretability in this domain. We propose to integrate the DAB perceptual dimensions in an interpretable layer of the DNNs and to jointly train the model using labels at the output layer for the task of interest (severity in our example) and labels for the interpretable layer. We limit the number of perceptual dimensions to the 4 broad dimensions listed above instead of the more elemental 38 perceptual dimensions.

3. Proposed model

Let us assume that there are n speech utterances with labels provided by clinicians at the speaker level. Those labels include a general severity rating (the network output) and ratings on the four perceptual dimensions we use to interpret the prediction of severity. After feature extraction we have a feature matrix, $\mathbf{X}$, with dimensions $n \times d$ where d is the dimension of the feature vector. We also have a scalar severity label vector $\mathbf{y}$ for each instance and a matrix interpretable layer label, $\mathbf{y}_\mathrm{I}$, with dimension $n \times 4$ representing the labels of the four perceptual dimensions.

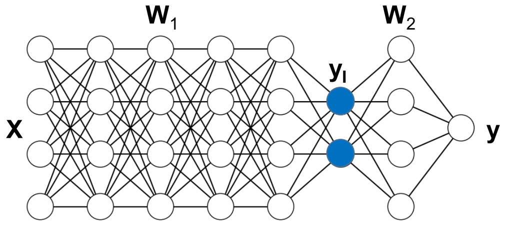

Figure 2: *The proposed interpretable DNN architecture for objective assessment of dysarthric speech. The layer with blue nodes is the interpretable Darley-Aronson-Brown layer.*

While the task of interest is severity prediction in this paper, this framework is easily extensible to other tasks (e.g. classifying based on disease instead of predicting severity).

3.1. DNNs architecture

The relationship between the input feature space and the final severity prediction is a complex one. For example, a dysarthric speaker can be severe because of atypical prosody or poor articulatory precision. DNNs are suitable to learn this complex mapping from the speech acoustics to the final prediction. To improve the interpretability of a notional DNN for predicting severity (e.g. see example in Fig. 1), we propose to add an intermediate layer before the final output as shown in Fig. 2. The different nodes in the DAB layer are trained to learn the labels for nasality, vocal quality, articulatory precision and prosody respectively. The DAB layer acts as an information-flow bottleneck and constrains the intermediate representation of the model in a way that clinicians can understand. This way, the end-user can backtrack from the model output by analyzing the output of the interpretable layer to gain additional insight into why the model made the final prediction that it did. In addition to providing an intermediate representation the clinician can understand, the new layer also acts as a regularizer that constrains the solution space of the DNN in order to prevent overfitting for small sample sizes.

3.2. DNNs training

In this section, we present two training strategies for our proposed model in Fig. 2. The two models either learn the weights of the network independently or jointly.

3.2.1. Sequential training

To train our proposed model, one strategy is to first train a network to predict the DAB representations with label $\mathbf{y}_\mathrm{I}$ and then train the remaining network to predict the final severity label with the output of the first network as the input of the second network. If we denote the weights prior to the interpretable layer by $\mathbf{W}_1$ and the weights following the interpretable layer by $\mathbf{W}_2$, then we define two cost functions and optimize them sequentially:

$$\mathbf{W}_1^* = \underset{\mathbf{W}_1}{\arg\min}\,\frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\mathbf{y}_\mathrm{I}^{(i)}}(\mathbf{W}_1) - \mathbf{y}_\mathrm{I}^{(i)}\right\|^2, \quad (1)$$

$$\mathbf{W}_2^* = \underset{\mathbf{W}_2}{\arg\min}\,\frac{1}{n}\sum_{i=1}^{n}(\hat{y}^{(i)}(\mathbf{W}_2) - y^{(i)})^2, \quad (2)$$

where $\hat{\mathbf{y}_\mathrm{I}^{(i)}}$ and $\hat{y}^{(i)}$ are the output of the interpretable layer and output layer of the i^{th} sample respectively.

3.2.2. *Joint training*

A different strategy is to concatenate the two networks and to jointly train both sets of model parameters, $\mathbf{W}_1$ and $\mathbf{W}_2$. The new objective function is

$$\mathbf{W}_1^*, \mathbf{W}_2^* = \underset{\mathbf{W}_1,\mathbf{W}_2}{\arg\min}(1-\lambda)\frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\mathbf{y}_\mathrm{I}^{(i)}}(\mathbf{W}_1) - \mathbf{y}_\mathrm{I}^{(i)}\right\|_2^2)+ \lambda\frac{1}{n}\sum_{i=1}^{n}(\hat{y}^{(i)}(\mathbf{W}_2)-y^{(i)})^2. \tag{3}$$

The parameter λ controls the tradeoff between the two parts of the objective function. If $\lambda = 1$, the proposed model does not consider interpretability and if $\lambda = 0$, the proposed model gives up on predicting severity and only focuses on predicting the four perceptual dimensions in the DAB layer. We vary λ to balance between model interpretability and the final prediction.

4. Experimental results

4.1. Dataset and feature extraction

Our data set was collected in the Motor Speech Disorders Lab at ASU. There are 87 speakers in this data set with four different dysarthria subtypes: ataxic dysarthria secondary to cerebellar degeneration (n = 16), mixed flaccid-spastic dysarthria secondary to amyotrophic lateral sclerosis (n = 15), hyperkinetic dysarthria secondary to Huntington's disease (n = 7), hypokinetic dysarthria secondary to Parkinson's disease (n = 41) and 8 healthy speakers. Each speaker read stimuli from visual prompts presented on a computer screen. Speech materials included 81 short phrases and 5 sentences [20, 21]. Fifteen master students from the ASU SLP program were asked to rate each subject on four perceptual dimensions: nasality, vocal quality, articulatory precision and prosody and then give a general impression of the severity of each patient (healthy speakers not included) based on their produced speech on a 1-7 (typical-severely atypical) scale. To integrate the ratings by multiple raters, we split the 15 raters into two sets - one set was used to train the model, the other set was used to test the model. For each of the two groups, we used the Evaluator Weighted Estimator (EWE) [22] to combine the multiple ratings into a single set of ratings by calculating the mean value weighted by individual reliability. The severity labels of healthy speakers were assigned as 1.

To increase the number of samples for DNNs training, we performed data augmentation on the original clean recorded speech materials. Two types of noise (meeting and office) [1] and two room impulse responses [2] were added to clean speech signals at two signal-to-noise ratios (5dB and 10dB) to simulate different environments. This results in over 50,000 speech utterances. The label scalar y and the label vector $\mathbf{y}_\mathrm{I}$ for each utterance were the severity label and labels of the four perceptual dimensions of the speaker who produced this utterance. Since the variability of utterance durations of each speaker is small, we consider this one to multiple label assignment strategy reliable.

Table 1: *Model performance for severity prediction. The best performance is denoted by bold numbers for each row.*

	Baseline	Sequential	Joint
PCC	0.821	0.811	**0.826**
SCC	0.807	0.799	**0.814**
MAE	0.729	0.709	**0.678**

Before feature extraction, speech samples were downsampled to 16kHz. Long-term speech features were extracted from the speech utterances, including: the envelope modulation spectrum [23], a representation of the slow amplitude modulations in a signal and the distribution of energy in the amplitude fluctuations across designated frequencies, captures rhythm information in the speech signal; The long-term average spectrum features and MFCC statistics [24] capture atypical average spectral information in the signal; Dysphonia features capture a patients' ability to control glottal movement; Correlation structure features [25, 6] that capture the evolution of vocal tract shape and dynamics at different time scale via auto- and cross- correlation analysis of formant tracks and MFCC. The feature dimension was 1201.

4.2. Model evaluation

We used speaker-level 5-fold cross validation (CV) to evaluate the performance of the different models. For each fold, 80% of the speakers were used for training the model and the remaining 20% speakers were used for evaluation. We further took 2000 utterances out from the training data as a validation set to monitor the training process and to tune the experimental settings. The severity prediction of the speakers in the evaluation set were calculated for only the clean speech samples. After predicting the severity of each speaker, we calculated the Pearson correlation coefficients (PCC, higher is better), Spearman correlation coefficients (SCC, higher is better) and the mean absolute error (MAE, lower is better) between the predicted ratings and ground-truth labels of dysarthric speakers. Tensorflow was used to construct and train the different DNNs architectures [26].

First, our baseline model was built using the architecture shown in Fig. 1. This model learns the direct mapping from the input acoustic features to severity without considering interpretability. This is a regression task and the cost function is the mean squared error (MSE) between the predicted severity rating and the ground-truth severity label. There were four hidden layers with 256 nodes per layer. The activation function was rectified linear unit (ReLU) for the hidden layers and linear for the output layer. The batch size was 256. Stochastic gradient descent (SGD) was used for optimization. The learning rate was set to 0.02 and exponentially decayed every 200 steps with a base of 0.8. The number of epochs was 30. All weight matrices were initialized using normal distribution with 0 mean and 0.01 standard deviation. Input acoustic features were zero-scored using the calculated mean and standard deviation from the training samples in each fold.

In our proposed model, we added the DAB layer before the output layer. Specifically, we replaced the output layer with the DAB layer. Then, we added one more hidden layer followed by the output layer as shown in Fig. 2. The network architecture before the DAB layer was identical to the baseline model. The number of nodes in the last hidden layer was set to 16. The activation function was ReLU for all hidden layers except for the

[1] http://parole.loria.fr/DEMAND/

[2] http://reverb2014.dereverberation.com/

Table 2: *Performance of the two training strategies of the interpretable model on the four perceptual dimensions in the DAB layer.*

	Nasality		Vocal quality		Articulatory Precision		Prosody	
	Sequential	Joint	Sequential	Joint	Sequential	Joint	Sequential	Joint
PCC	0.739	0.749	0.705	0.735	0.808	0.816	0.781	0.788
SCC	0.708	0.727	0.684	0.709	0.778	0.786	0.756	0.771
MAE	1.149	1.149	0.823	0.798	0.782	0.748	0.963	0.942

DAB layer and the output layer (linear for both). As mentioned in section 3, the first strategy to train this model was sequential training. For training the parameters before the DAB layer, we used the same experimental settings as the baseline model. After training, the model parameters were saved and fixed. During the 2nd training phase, only the remaining network parameters (those after the interpretable layer) were updated. In the 2nd training phase, we used a smaller learning rate (0.001 with the same decay rate) and only ran 10 epochs.

Next, we also validated our proposed model using the joint training procedure, in which all network parameters were jointly optimized. Experimental settings were the same with the baseline model. The only difference was that the objective function here had two parts as shown in eqn. 3. We varied λ from 0.1 to 0.9 with a step size of 0.2 to empirically evaluate the model performance as a function of λ. We skipped the two extreme cases ($\lambda = 0$ and $\lambda = 1$) because either predictive ability or model interpretability was lost.

4.3. Results analysis

Note that all of the results are the average of five Monte Carlo trials. We first analyze at the performance of three different methods that predict the severity of each speaker. The three measurements of the final severity prediction using three different methods are shown in table 1. The results of joint training are presented for the optimal λ. As the table shows, we see that the sequential training procedure of the interpretable DNN results in a slightly lower PCC and SCC when compared against the baseline; but the MAE is also lower. As expected, the joint training strategy yields the best performance in terms of all three measurements and provides the best predictive ability.

In addition to predicting the final severity output, we also ensure that the proposed model yields reasonable interpretations by providing good predictions of the other four perceptual dimensions used in the DAB layer of the network. In table 2, we use the same metrics to evaluate the proposed models on the four perceptual dimensions. The baseline model is not included since it does not provide a prediction for these dimensions. The results of joint training are shown for the optimal λ. As before, the joint training procedure results in reduced MAE and increased PCC and SCC in general. A possible reason for this is that the joint prediction has the benefit of regularizing the model in a manner similar to transfer learning, or multi-task learning. It is clear that the perceptual dimensions are somehow related to the final severity rating. By forcing the model to learn the dimensions and final severity rating together, this constrains the solution space of the parameter set and the result is a lower overall error. As the table shows, the nasality dimension provides the highest MAE; however this is also the most difficult dimension to label by experts as well.

Next, we show how the performance of joint training model varies with λ in Fig. 3. Here we only show the severity prediction results since the interpretation dimensions follow a similar trend. Along the x-axis, we vary lambda from 0.1 to 0.9 with a step size of 0.2. The y-axis on the left is the MAE and the y-axis on the right is the PCC and SCC. We find that the highest performance is achieved for $\lambda = 0.7$. Increasing λ past 0.7 results in a rapid decline in performance since, when $\lambda \approx 1$, the supervision information in the DAB layer provides little regularization and the interpretable layer does not accurately model the DAB dimensions. Correlation-based measurements follow the same trend as the MAE.

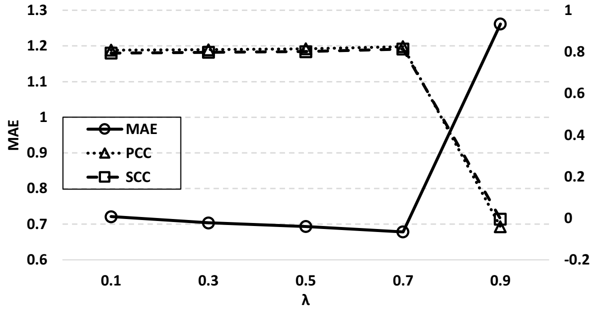

Figure 3: *Performance of the joint training DNN for varying* λ.

5. Conclusion

In this paper we propose an interpretable objective severity assessment algorithm of dysarthric speech based on DNNs. An intermediate DAB layer with a representation understood by speech language pathologists and neurologists is added to the DNN. The model is trained with a scalar severity label at the output of the network and intermediate labels that describe how atypical the speech is along four perceptual dimensions in the DAB layer. We investigate two strategies to train the proposed model: sequential training and joint training. We compare the proposed model with a baseline DNN that does not account for model interpretability. Experimental results demonstrate that using the proposed joint training our model can both provide better prediction accuracy and interpretability compared to sequential training and a baseline model.

In future, we will explore additional prediction tasks beyond severity estimation. For example, disease diagnosis based on speech analytics is an area of great interest currently; it would be useful for a diagnostic DNN to also provide an intermediate representation (perhaps one based on the DAB dimensions) so that clinicians can gain additional insight into why the DNN made the decision that it did. For more complicated tasks, additional interpretable nodes in the DAB layer can be added to improve both model performance and interpretability.

6. Acknowledgements

The authors gratefully acknowledge partial support of this work by an NSF STTR Phase I grant awarded to Aural Analytics, LLC and an NIH-NIDCD Grant 1R21DC013812.

7. References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144.

[2] Z. C. Lipton, "The mythos of model interpretability," *arXiv preprint arXiv:1606.03490*, 2016.

[3] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, 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*, 2015.

[4] J. Stone, A. Carson, and M. Sharpe, "Functional symptoms and signs in neurology: assessment and diagnosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 76, no. suppl 1, pp. i2–i12, 2005.

[5] G. An, D. G. Brizan, M. Ma, M. Morales, A. R. Syed, and A. Rosenberg, "Automatic recognition of unified parkinson's disease rating from speech with acoustic, i-vector and phonotactic features," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] J. R. Williamson, T. F. Quatieri, B. S. Helfer, J. Perricone, S. S. Ghosh, G. Ciccarelli, and D. D. Mehta, "Segment-dependent dynamics in predicting parkinson's disease," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

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

[8] M. Tu, Y. Jiao, V. Berisha, and J. M. Liss, "Models for objective evaluation of dysarthric speech from data annotated by multiple listeners," in *Signals, Systems and Computers, 2016 50th Asilomar Conference on*. IEEE, 2016, pp. 827–830.

[9] M. Tu, V. Berisha, and J. Liss, "Objective assessment of pathological speech using distribution regression," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5565–5569.

[10] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[11] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *European Conference on Computer Vision*. Springer, 2016, pp. 3–19.

[12] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: understanding rating dimensions with review text," in *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 2013, pp. 165–172.

[13] M. Tu, A. Wisler, V. Berisha, and J. M. Liss, "The relationship between perceptual disturbances in dysarthric speech and automatic speech recognition performance," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. EL416–EL422, 2016.

[14] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 12, no. 2, pp. 246–269, 1969.

[15] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks." in *Interspeech*, vol. 237, 2011, p. 240.

[16] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Joint training of frontend and back-end deep neural networks for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4375–4379.

[17] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific*. IEEE, 2015, pp. 1225–1237.

[18] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1717–1724.

[19] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu, "Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2740–2748.

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

[21] J. M. Liss, L. White, S. L. Mattys, K. Lansford, A. J. Lotto, S. M. Spitzer, and J. N. Caviness, "Quantifying speech rhythm abnormalities in the dysarthrias," *Journal of Speech, Language, and Hearing Research*, vol. 52, no. 5, pp. 1334–1352, 2009.

[22] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10, pp. 787–800, 2007.

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

[24] P. Rose, *Forensic speaker identification*. CRC Press, 2003.

[25] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. 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.

[26] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.