Dysarthria Intelligibility Assessment in a Factor Analysis Total Variability Space

David Martínez[1], *Phil Green*[2], *Heidi Christensen*[2]

[1]Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain
[2]Computer Science, University of Sheffield, United Kingdom
david@unizar.es, (p.green | h.christensen)@dcs.shef.ac.uk

Abstract

Speech technologies are more important every day to assist people with speech disorders. They can help to increase their quality of life or help clinicians to make a diagnosis. In this paper a new methodology based on a total variability subspace modelled by factor analysis is proposed to assess the intelligibility of people with dysarthria. The acoustic information of each recording is efficiently compressed and a Pearson correlation of 0.91 between the vectors in this subspace (iVectors) and the intelligibility is obtained. As acoustic information only perceptual linear prediction features are used. The experiments are conducted on Universal Access Speech database. Also a new error metric to overcome the subjectivity in the intelligibility labels is proposed.

Index Terms: dysarthria, intelligibility assessment, iVectors, factor analysis, prediction

1. Introduction

The term dysarthria is used to refer to any of the speech disorders that are due to disturbances in neuromuscular control of the speech mechanism resulting from impairment of any of the basic motor processes involved in speech production [1]. This can affect respiration, phonation, resonance, articulation, and prosody, and can provoke abnormal characteristics in speech quality and reduced intelligibility. Six major types of dysarthria can be found depending on the affected area of the neuromotor system: flaccid associated with lower motor neurons, spastic with upper motor neurons linked to the cerebral cortex, ataxic with the cerebellum, hyperkinetic and hypokinetic both with the extrapyramidal system, and mixed which affects more than one of the previous areas [2].

Up to now clinical diagnoses of dysarthric speakers have been conducted by speech therapists, which means that there is a subjective contribution in the evaluations, resulting in disagreements among experts. In order to remove as much as possible this subjectivity, standard methods to assess dysarthria diagnosis have been developed, like the *Dysarthria Profile* [3], the *Frenchay Dysarthria Assessment* (FDA) [4], or the *Dysarthria Examination Battery* (DEB) [5]. All of them contain a section dedicated to rate intelligibility, because the level of intelligibility is an indication of the type of dysarthria, of the degree of the disorder, and of the relative contribution of the basic physiological mechanisms [6]. One of the benefits that speech technology brings to speech therapists is the objectivity and replication of the results, and consequently, some of these tests have introduced this type of technology in their evaluations. For example, in [7] an automatic speech recognition (ASR) system is used to rate intelligibility in a computerized version of the FDA.

Several techniques have been tested for predicting intelligibility of dysarthric speakers. Basically, two main approaches are found in the bibliography. In the first, the speech intelligibility is calculated directly from the word accuracy (WA) obtained from an ASR system. That is, it is considered that intelligible speech will obtain low word error rate (WER) on an ASR system trained on clean and presumably highly intelligible speech, and low intelligible speech will obtain high WER [8, 9, 10, 11]. One of the main criticisms of these systems is that they are trained only on non-dysarthric speakers and the result can be unpredictable for very severe subjects [12]. In the second, different features are extracted from speech and used to build an intelligibility predictor model [13, 14, 15, 16, 17, 18]. The use of an ASR or automatic speech alignment (ASA) system is restricted to feature extraction [12, 19].

In this paper, it is proposed to predict intelligibility from the vectors that represent each utterance in the acoustic space, in a similar way to the work in [16]. However, instead of using GMM-based supervectors as they do, points in the total variability subspace, or *iVectors*, a state-of-the-art approach succesfully applied in the field of speaker recognition [20], are used. The total variability subspace is a low-dimension subspace where the main variabilities describing the data are kept. Our assumption is that those dimensions contain also information about intelligibility, as is confirmed in the experimental section of the work.

One of the main problems of working with dysarthric speech is the scarcity of data within the available databases [21]. The data recording requires several repetitions of words involving difficult movements of the speech articulators, which can be very exhausting for some dysarthric conditions. In this paper we work on the Universal Access Speech (UAS) database [22]. In this database 15 dysarthric speakers are available with different degrees of intelligibility. Given the limited number of speakers, the experiments conducted in previous studies on this database [13, 14, 15, 11], are driven in a speaker-dependent manner, where the speaker under test is also in the training dataset, but the intelligibility is evaluated for utterances not seen during the training. In this work we follow the same speaker-dependent strategy. Note that the real application of this task is limited to small scenarios where all speakers are known beforehand. However, this is the case in many projects in the real world, like *homeService* [1], the one in which this work is framed.

The paper is organized as follows: in Section 2 the databases used for the experiments are presented; in Section 3 the system architecture is described; in Section 4 the experiments are detailed; and in Section 5 the conclusions of the work are drawn.

[1]http://www.natural-speech-technology.org/homeService

Speaker Label	Age	Speech Intelligibility (%)	Dysarthria Diagnosis
M01	>18	very low (15%)	Spastic
M04	>18	very low (2%)	Spastic
M05	21	mid (58%)	Spastic
M07	58	low (28%)	Spastic
M08	28	high (93%)	Spastic
M09	18	high (86%)	Spastic
M10	21	high (93%)	Mixed
M11	48	mid (62%)	Athetoid
M12	19	very low (7.4 %)	Mixed
M14	40	high (90.4%)	Spastic
M16	-	low (43%)	Spastic
F02	30	low (29%)	Spastic
F03	51	very low (6%)	Spastic
F04	18	mid (62%)	Athetoid
F05	22	high (95%)	Spastic

Table 1: *UAS speakers information*

2. Audio Material

Two databases are used in the training process: UAS and Wall Street Journal Database 1 (WSJ1) [23]. The sampling rate is fixed at 16 kHz in both. Next we describe them and explain how they are used in this study.

2.1. Universal Access Speech Database

This is a dysarthric speech database recorded from 19 speakers with cerebral palsy. We have available data from 15 of them. Data were recorded in an 8-microphone array at 48 kHz and 1 digital video camera. For each speaker 765 words were recorded in 3 blocks of 255, 155 of which are common to the 3 blocks and 100 are uncommon words that differ across them. The 155-word blocks include 10 digits, 26 radio alphabet letters, 19 computer commands, and the 100 most common words in the Brown corpus of written English. To calculate the intelligibility rate of each speaker five naive listeners were asked to provide ortographic transcriptions of each word. The correct percentages for each speaker obtained by the five listeners were averaged to calculate the speaker's intelligibility. In table 1 a summary of each speaker in the database and their intelligibilities can be seen. For more information about the database, please refer to [22].

For our experiments, only microphone 6 is used, dividing the data in two, training and test. For testing all uncommon words are reserved, and for training the rest. The training part is used to train the predictor explained in Section 3.4. This is a similar configuration to the one in [13].

2.2. Wall Street Journal 1

This is a general-purpose English, large vocabulary, natural language, high perplexity corpus containing a substantial quantity of speech data (77800 training utterances including about 73 hours of speech). It includes read speech and spontaneous dictation by journalists. The database also contains development and test datasets in a "Hub and Spoke" paradigm to probe specific areas of interest. Each of them contains 7500 waveforms, about 11 hours of speech. Data were collected using two microphones at a sampling rate of 16 kHz. For more information please consult [23].

This database is selected because it contains a large amount of speech in American English, like UAS, so we are able to train our maximum likelihood (ML) models described in next section, the Gaussian mixture model (GMM) and factor analysis (FA) front-end, more reliably than using only UAS, as shown in Section 4.1. Also because both databases mainly contain read speech (spoken to prompts in UAS). Only the following directories of the database are used, which are basically those including clean speech (for a complete overview of the database directory system, please visit [2]):

- si_tr_s (SI, training, 150 WSJ sentences)
- si_tr_l (SI, training, 1200 WSJ sentences)
- si_tr_j (SI, training, journalists, 200 WSJ sentences)
- si_tr_jd (SI, training, journalists, spon. dictation)
- si_dt_20 (Hub 1 test data)
- si_dt_05 (Hub 2 test data)
- si_dt_jd (Spoke 9 test data)
- si_dt_s1 (Spoke 1 test data)
- si_dt_s2/sjm (Spoke 2 test data)
- si_dt_s4/inc_adp (Spoke 4 test data)

[2] http://www.ldc.upenn.edu/Catalog/docs/LDC94S13A/wsj1.txt

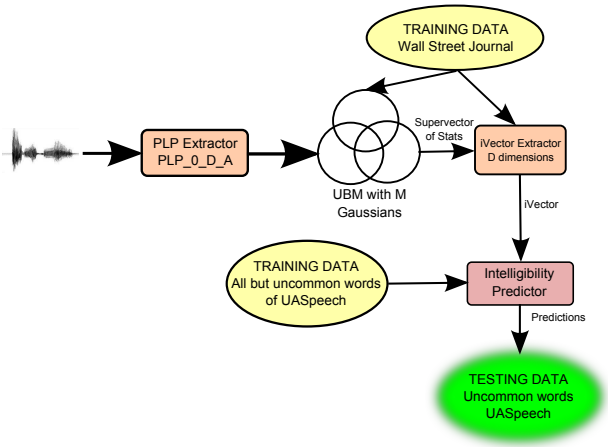

Figure 1: System Architecture

3. System Architecture

The system architecture is depicted in figure 1. The idea of this system is to make intelligibility predictions from the vectors lying in the the total variability space, a low-dimension subspace obtained from an FA model. These vectors, also known as iVectors, have the advantages of being fixed-length and low-dimension. Each one represents an utterance, then predictions can be made directly from them. In next subsections, every component of the system is explained.

3.1. Acoustic Features

The input to our system are 12 perceptual linear prediction features (PLP) [24] plus energy, with derivatives and accelerations, to obtain a 39 dimension vector every 10 ms, in 25 ms length windows. These features use 3 findings of the pshychophysics of hearing: the critical band spectral resolution, the equal-loudness curve, and the intensity-loudness power law. Based on this, our hypothesis is that they carry information of intelligibility.

3.2. Gaussian Mixture Model and Sufficient Statistics

A GMM [25] is a multimodal distribution typically used in speech processing. A universal GMM of our data is created by running 20 iterations of the EM algorithm. Once the model is trained, zeroth (N) and first (F) order Baum-Welch statistics for each utterance are obtained as:

$$N_c = \sum_{t=1}^{L} \mathrm{P}(c|x_t, \Omega), \tag{1}$$

$$F_c = \sum_{t=1}^{L} \mathrm{P}(c|x_t, \Omega) x_t, \tag{2}$$

where L is the number of frames in a given file, and $P(c|x_t, \Omega)$ is the posterior probability of mixture component c generating the PLP vector x_t, for a model with parameters Ω and C components. The stats supervector is built by concatenating the stats for each Gaussian component. This is the input for the iVector extractor, described in the following subsection. Note that in [16] they use these supervectors to predict intelligibility, whereas we use iVectors extracted with an FA model instead.

3.3. Factor Analsysis Front-end

FA assumes that the main variability of the signal lies in a low-dimension subspace. Thus the mean supervector of our previous GMM is modelled as

$$m(s) = m_0 + Ti(s), \tag{3}$$

where m_0 is the mean supervector of the GMM, i is the iVector that has an a priori standard normal distribution $\mathcal{N}(0, 1)$, and T is a KxD matrix which translates the iVectors from their low-dimension total variability space to the high-dimension space where the model $m(s)$ lies, being D the iVector dimension and $K = 39$xC the dimension of the supervectors. The training of T is done by alternating an ML step with a minimum divergence step (MD). The iVector is completely characterized by its posterior distribution conditioned to the sufficient statistics, and follow a Gaussian distribution [20].

The T matrix spans a space trained on the main variabilities of the signal, and our assumption is that those variabilities contain information about the intelligibility. Note that one iVector is obtained for each utterance, that is used to predict the intelligibility.

3.4. Predictor

Two different predictors have been investigated, a linear predictor and a support vector regression predictor (SVR). In the first, the sum of squares error function

$$E(w) = \frac{1}{2}\sum_{r=1}^{R}(y_r - \mathbf{w}^t\phi(\mathbf{i_r}))^2, \tag{4}$$

is minimized to obtain a linear function of the iVectors in the form $\mathbf{w} \cdot \phi(i(s))$ [26], that converts the iVector into an intelligibility rate. R is the number of utterances for training, y_r is the target intelligibility value for utterance r, $\mathbf{w}$ is a vector of weights of dimension D+1, and $\phi(\mathbf{i_r})$ is a vector of D+1 dimension basis functions, in our case $\phi(i_j) = i_j$ and $\phi_0 = 1$.

In SVR, the number of errors is minimized subject to the constraint that only a set of vectors which are not further than a margin ϵ from the regression curve, called support vectors, are used for training [27]. Two flavours of SVR are tested, ϵ-SVR, where the margin ϵ is directly selected, and ν-SVR, where the margin ϵ is embedded into the cost function and allows controlling the number of support vector and training errors.

4. Experiments

This section is divided into two subsections, one for the linear predictor and one for the SVR predictor. In the first the optimal number of Gaussians of the GMM and optimal dimension for iVectors is selected. Then these values are used for the SVR and a comparison is made between the two flavours of SVR. Results are given in terms of Pearson correlation (PC), Spearman correlation (SC), and root mean square error (RMSE) between the rated and predicted intelligibilities divided by 100 (intelligibilities $\in$ [0,1]). Finally, a new metric is proposed to overcome the intrinsic variability that intelligibility has by nature.

Nr. Gaussians	512	1024							2048
iVector Dim	400	5	10	50	100	200	400	600	400
PC	0.8739	0.7623	0.8132	0.8512	0.8602	0.8718	0.8806	**0.8813**	0.8779
SC	0.8652	0.7605	0.8209	0.8493	0.8592	0.8659	0.8718	**0.8724**	**0.8724**
RMSE	0.1655	0.2230	0.1997	0.1784	0.1731	0.1665	0.1614	**0.1609**	0.1627

Table 2: *Results with the Linear Predictor*

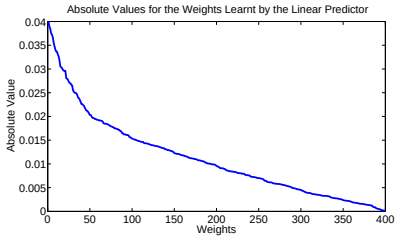

Figure 2: Weights of the linear predictor sorted in descending order in the x-axis according to its absolute value in the y-axis

4.1. Linear Prediction

In Table 2 it can be seen that the correlation between iVectors and intelligibility increases as we increase the number of Gaussians in the GMM and the iVector dimensionality. The best results are obtained for 1024 Gaussians and 600 dimensions, and increasing the number of Gaussians does not give any further improvement.

It is remarkable that even a low iVector dimensionality like 10 still gives PC over 0.8, which indicates that most of the intelligibility information is contained in a few dimensions. Inspection of the ranked absolute value of the weights $|\mathbf{w}|$ for 400 dimension iVectors and a GMM with 1024 components in figure 2 reveals that after 50 dimensions the weight value has been halved, and the weight 350 is about a tenth of the first. This explains why adding more than 400 dimensions gives so small improvements.

In order to check if some improvement can be gained from using all available data in UAS, the predictor is trained with all microphones and the test is made again over microphone 6 only. This gives a PC=0.8918, a SC=0.8789, and a RMSE=0.1549, for the system with 1024 Gaussians and 400 dimension iVectors. Note that the improvement comes from having the same training data but recorded in different microphones. Since it is not a huge improvement, for the next experiments we continue working only with microphone 6 for both training and testing.

One possible critcism of these experiments is that the training of the GMM and FA front-end is made with WSJ1, a database very different to UAS. One might think that this kind of training could create unpredictable iVectors for the dysarthric speech. To check if this happens the same experiment shown in table 2 is made but using only UAS for all training. In table 3 these results are shown. It can be seen that the results are much worse, indicating that having more data for training the ML models helps, even if these data are not of dysarthric speakers.

4.2. Support Vector Regression Prediction

The first experiment with SVR aims at checking if ϵ-SVR or ν-SVR give better accuracy than the linear predictor, and which of both performs better. The system with 1024 Gaussians and 400 dimension iVectors is selected for the rest of the experiments with SVR, since 600 dimensions does not give great gains and

Number of Gaussians	512			1024		
Dimension of iVectors	50	100	400	50	100	400
Pearson Correlation	0.4705	0.5135	0.5949	0.2777	0.4009	0.4913
Spearman Correlation	0.4907	0.5337	0.6125	0.2894	0.4210	0.5100
RMSE	0.2993	0.2917	0.2728	0.3258	0.3120	0.2995

Table 3: *Results with the Linear Predictor, GMM and FA models also trained on UAS*

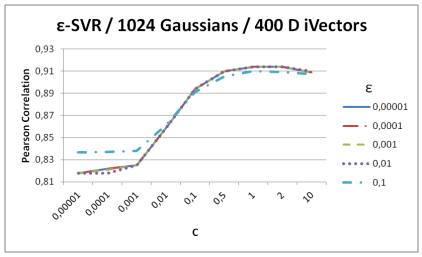

Figure 3: PC with ϵ-SVR, 1024 Gaussians and 400 dimension iVectors

is computationally quite more expensive. In figures 3 and 4, PC is plotted as a function of C, the regularization constant of SVR, for different values of ϵ and ν. The optimal C is equal to 1 in both cases, the optimal ϵ is 0.01, and the optimal ν is 1.

The best results are obtained for ν-SVR with $C = 1$ and $\nu = 1$, where PC=0.9137, SC=0.9004, and RMSE=0.1400. For comparison, in [13] a PC=0.94, a SC=0.89, and a RMSE=0.1860 are obtained, using only ten spastic speakers from the UAS database, and a system predicting from a set of six features representing atypical vocal source excitation, temporal dynamics, and prosody. In our system many more dimensions are used but they are all derived from PLP features. In figure 5 the mean and standard deviation obtained for each speaker in the database are shown. The means follow pretty well a regression line with slope 1, which is confirmed by PC and SC, but the standard deviations can look high in some cases. We should note that, first an intelligibility measurement has an intrinsic subjective component and it is not realistic to give a unique intelligibility rate for each speaker, since different listeners perceive different intelligibilities in one utterance; and second, each dysarthric speaker is assigned a unique intelligibility rate, but every time he/she pronounces a different sentence, the intelligibility can change, and again, it is not real to label each speaker with a unique intelligibility. Consequently, it is normal to observe a relatively high standard deviation.

Intelligibility is a subjective feature and there should not be gold standard measurements of it. It has an intrinsic variability among listeners, and to reduce this effect a new metric to evaluate systems performance is proposed. In [22], 4 groups are defined according to the intelligibility given by the listeners: very low (0%-25%), low (26%-50%), high (51%-75%), and very high (76%-100%). Thus the problem could be also treated as a classification problem to decide which group a speech utterance belongs to. However, this would lose the continuity of intelligibility, and for instance, a sample labelled as low, and predicted as high with a 51% of intelligibility would count as an error. What we propose is that an utterance counts as an error only if it is over or under the target intelligibility by a specific absolute percentage. Following the division commented before, each covering a 25% of the intelligibility, we propose to record an error if our prediction is an absolute 12.5% larger or smaller than the labelled intelligibility. For instance, for speaker M05 who has an intelligibility of 58%, any value over 70.5% or under 45.5% would be considered as an error. In the same manner, for speaker M12 who has an intelligibility of 7.4%, only values over 19.9% would be considered as an error, since we can not obtain intelligibilities under 0%. In this way, we are more flexible with the intelligibility decision, and consider a wider range of intelligibilities as correct to absorb the intrinsic variability that intelligibility has. This error rate would be:

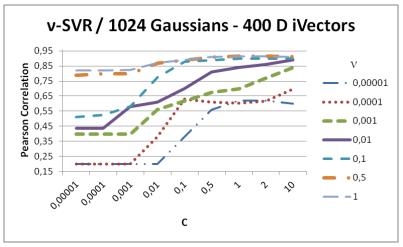

Figure 4: PC with ν-SVR, 1024 Gaussians and 400 dimension iVectors

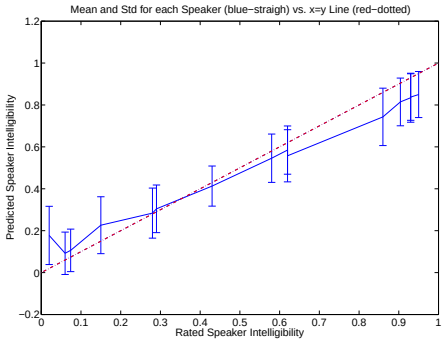

Figure 5: Mean and standard deviation of predicted intelligibilities per speaker (blue straight line) in system with 1024 Gaussians and 400 dimension iVectors, and ν-SVR with C=1 and ν=1, compared with a slope 1 line (red dotted line)

$$error\_rate_{12.5\%} = \frac{C^+ + C^-}{N_t}, \quad (5)$$

$C^+ = \sum(predicted\_values > target\_value + 12.5\%)$,
$C^- = \sum(predicted\_values < target\_value - 12.5\%)$,
N_t = number of test utterances.

For ν-SVR prediction, C=1, ν=1, 1024 Gaussians and 400 dimension iVectors, $error\_rate_{12.5\%} = 0.3266$.

5. Conclusions

The total variability or iVector subspace recently used in speaker verification is successfully applied to dysarthric speech intelligibility assessment. This subspace retains in a few dimensions most of the variability of the signal, and our assumption stating that that variability correlates well with the speaker intelligibility is shown. Linear, ϵ- and ν-SVR prediction are tested and a PC=0.9137, SC=0.9004, and RMSE=0.1400 are obtained for the last, using a system with a GMM with 1024 components, 400 dimension iVectors, ν=1, and C=1. As input features, only PLP coefficientes are used, and further improvements are expected if features specifically designed for intelligibility assessment were used. Finally, a new error metric designed to overcome the intrinsic subjectivity of intelligbility reflected in the labels is presented. In this metric only the files whose prediction is a 12.5% higher or lower than the target intelligibility are considered as errors.

6. Acknowledgements

This work has been done during an internship of David Martínez in SPandH group of the University of Sheffield, funded by the Spanish Government and European Union (FEDER) under project TIN2011-28169-C05-02.

7. References

[1] F. Darley, E. Aronson, and R. Brown, *Motor Speech Disorders*. Philadelphia: WB Saunders, 1975.

[2] P. Enderby, "Disorders of Communication: Dysarthria," in *Handbook of Clinical Neurology*, 110th ed., M. Barnes, D. Good, M. Aminoff, F. Boller, and D. Swaab, Eds. Elsevier B.V., 2013, ch. 22, pp. 273–281.

[3] S. Robertson, *Dysarthria Profile*. Winslow Press, 1982.

[4] P. Enderby, *Frenchay Dysarthria Assessment*. Austin, Texas: Pro-Ed, 1983.

[5] S. Drummond, *Dysarthria Examination Battery*. Communication Skill Builders, 1993.

[6] E. A. Strand, "Dysarthrias: Management," in *The MIT Encyclopedia of Communication Disorders*, R. D. Kent, Ed., 2004, pp. 129–132.

[7] J. Carmichael, "Introducing Objective Acoustic Metrics for the Frenchay Dysarthria Assessment Procedure," Ph.D. dissertation, University of Sheffield, 2007.

[8] P. Doyle, H. Leeper, A. Kotler, N. Thomas-Stonell, C. O'Neill, M. Dylke, and K. Rolls, "Dysarthric Speech: a Comparison of Computerized Speech Recognition and Listener Intelligibility," *Journal of Rehabilitation Research and Development*, vol. 34, no. 3, pp. 309–316, Jul. 1997.

[9] J. Carmichael and P. Green, "Revisiting Dysarthria Assessment Intelligibility Metrics," in *Proceedings of the 8th International Conference on Spoken Language Processing*, Jeju Island, Korea, 2004.

[10] H. Sharma, M. Hasegawa-Johnson, J. Gunderson, and A. Perlman, "Universal Access: Preliminary Experiments in Dysarthric Speech Recognition," in *Interspeech*, Brighton, UK, 2009, pp. 7–10.

[11] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A Comparative Study of Adaptive, Automatic Recognition of Disordered Speech," in *Interspeech*, Portland, OR, 2012.

[12] C. Middag, J.-P. Martens, G. Van Nuffelen, and M. De Bodt, "Automated Intelligibility Assessment of Pathological Speech Using Phonological Features," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, 2009.

[13] T. Falk, W. Chan, and F. Shein, "Characterization of Atypical Vocal Source Excitation, Temporal Dynamics and Prosody for Objective Measurement of Dysarthric Word Intelligibility," *Speech Communication*, vol. 54, no. 5, pp. 622–631, Jun. 2012.

[14] T. Falk, R. Hummel, and W. Chan, "Quantifying Perturbations in Temporal Dynamics for Automated Assessment of Spastic Dysarthric Speech Intelligibility," *Intl. Conference on Acoustics, Speech and Signal Processing*, pp. 4480–4483, 2011.

[15] M. Paja and T. Falk, "Automated Dysarthria Severity Classification for Improved Objective Intelligibility Assessment of Spastic Dysarthric Speech," in *Interspeech*, Portland, OR, 2012.

[16] T. Bocklet, K. Riedhammer, E. Nöth, U. Eysholdt, and T. Haderlein, "Automatic Intelligibility Assessment of Speakers After Laryngeal Cancer by Means of Acoustic Modeling," *Journal of Voice*, vol. 26, no. 3, pp. 390–397, May 2012.

[17] M. De Bodt, M. Hernández-Díaz Huici, and P. Van De Heyning, "Intelligibility as a Linear Combination of Dimensions in Dysarthric Speech," *Journal of Communication Disorders*, vol. 35, no. 3, pp. 283–292, 2002.

[18] C. Middag, T. Bocklet, J.-P. Martens, and N. Elmar, "Combining Phonological and Acoustic ASR-Free Features for Pathological Speech Itelligibility Assessment," in *Interspeech*, Florence, Italy, 2011.

[19] G. Van Nuffelen, C. Middag, M. D. Bodt, and J. Martens, "Speech Technology-Based Assessment of Phoneme Intelligibility in Dysarthria," *International Journal of Language & Communication Disorders*, vol. 44, no. 5, pp. 716–730, 2009.

[20] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[21] P. Green, J. Carmichael, and A. Hatzis, "Automatic Speech Recognition with Sparse Training Data for Dysarthric Speakers," in *Proc. 8th European Conference on Speech Communication Technology (Eurospeech)*, Geneva, Switzerland, 2003.

[22] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric Speech Database for Universal Access Research," in *Interspeech*, Brisbane, Australia, 2008, pp. 1741–1744.

[23] D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," in *HLT'91 Proc. of the Workshop on Speech and Natural Language*, Pacific Grove, CA, 1991.

[24] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *Journal Acoustic Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[25] D. Reynolds and R. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, 1995.

[26] C. Bishop, *Pattern Recognition and Machine Learning*, M. Jordan, J. Kleinberg, and B. Schölkopf, Eds. Springer, 2006.

[27] A. Smola and B. Schölkopf, "A Tutorial on Support Vector Regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004.