Intelligibility Assessment and Speech Recognizer Word Accuracy Rate Prediction for Dysarthric Speakers in a Factor Analysis Subspace

DAVID MARTÍNEZ and EDUARDO LLEIDA, Universidad de Zaragoza
PHIL GREEN and HEIDI CHRISTENSEN, University of Sheffield
ALFONSO ORTEGA and ANTONIO MIGUEL, Universidad de Zaragoza

Automated intelligibility assessments can support speech and language therapists in determining the type of dysarthria presented by their clients. Such assessments can also help predict how well a person with dysarthria might cope with a voice interface to assistive technology. Our approach to intelligibility assessment is based on *iVectors*, a set of measures that capture many aspects of a person's speech, including intelligibility. The major advantage of *iVectors* is that they compress all acoustic information contained in an utterance into a reduced number of measures, and they are very suitable to be used with simple predictors. We show that intelligibility assessments work best if there is a pre-existing set of words annotated for intelligibility from the speaker to be evaluated, which can be used for training our system. We discuss the implications of our findings for practice.

Categories and Subject Descriptors: I.5.2 [**Design Methodology**]: Classifier Design and Evaluation

General Terms: Algorithms, Design, Performance

Additional Key Words and Phrases: Intelligibility assessment, ASR accuracy prediction, dysarthria, factor analysis, iVectors

ACM Reference Format:
David Martínez, Eduardo Lleida, Phil Green, Heidi Christensen, Alfonso Ortega, and Antonio Miguel. 2015. Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace. ACM Trans. Access. Comput. 6, 3, Article 10 (May 2015), 21 pages.
DOI: http://dx.doi.org/10.1145/2746405

1. INTRODUCTION

The term *dysarthria* is used to refer to any of the speech disorders caused by disturbances in neuromuscular control of the speech mechanism and resulting from impairment of any of the basic motor processes involved in speech production [Darley et al. 1975]. 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

This work was funded by the Spanish government through project TIN2011-28169-C05-02, and by the European Union through projects INNPACTO IPT-2011-1696-390000 (FEDER) and Iris, which has received funding from the Seventh Framework Programme for research, technological development, and demonstration under grant agreement 610986.
Authors' addresses: D. Martínez, E. Lleida, A. Ortega, and A. Miguel, Departamento Ingeniería Electrónica y Comunicaciones, Universidad de Zaragoza, María de Luna, 1, Zaragoza 50018, Spain; emails: {david, lleida, ortega, amiguel}@unizar.es; P. Green and H. Christensen, Dept of Computer Science, University of Sheffield, Regent Court, Portobello, Sheffield S1 4DP, UK; emails: p.green@shef.ac.uk, heidi.christensen@sheffield.ac.uk.

DOI: http://dx.doi.org/10.1145/2746405

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 [Enderby 2013].

Clinical diagnoses of dysarthric speakers have been traditionally conducted by speech therapists, which means that there is a subjective contribution in the evaluations, resulting in disagreements among experts. To remove as much of this subjectivity as possible, standard methods to assess dysarthria diagnosis have been developed, like the *Dysarthria Profile* [Robertson 1982], the *Frenchay Dysarthria Assessment* (FDA) [Enderby 1983], and the *Dysarthria Examination Battery* (DEB) [Drummond 1993]. All of these contain a section dedicated to rating intelligibility, because the level of intelligibility is an indication of the type of dysarthria, degree of the disorder, and relative contribution of the basic physiological mechanisms [Strand 2004]. One of the benefits that speech technology brings to speech therapists is the objectivity and replication of the results; consequently, some implementations of these tests have introduced this type of technology. For example, in Carmichael [2007], an automatic speech recognition (ASR) system is used to rate intelligibility in a computerized version of the FDA.

Basically, two main approaches are found in the literature for predicting intelligibility of dysarthric speakers. In the first approach, the speech intelligibility is calculated directly from the word accuracy rate (*Accuracy*) obtained from an ASR system. This is based on the observation that intelligible speech will obtain high *Accuracy* on an ASR system trained on typical and presumably highly intelligible speech, and low intelligible speech will obtain low *Accuracy* [Doyle et al. 1997; Carmichael and Green 2004; Sharma et al. 2009; Christensen et al. 2012]. One of the main weaknesses of these systems is that they are trained only on nondysarthric speakers, and the result can be unpredictable for very severe subjects [Middag et al. 2009]. In the second approach, different features are extracted from speech and used to build an intelligibility assessment system [Middag et al. 2011; Falk et al. 2011; Bocklet et al. 2012; Falk et al. 2012; Paja and Falk 2012]. These experiments are supported by perceptual studies that show how intelligibility can be expressed as a linear function of multiple speech dimensions [De Bodt et al. 2002]. In this approach, the use of a speech recognizer or an automatic speech alignment (ASA) system is restricted to feature extraction [Van Neuffelen et al. 2009; Middag et al. 2009].

In this article, the information contained in a whole utterance (in our case, each utterance contains a single word) is compressed and represented as points in the total variability subspace, or *iVectors*, a state-of-the-art approach successfully applied in the field of speaker recognition [Dehak et al. 2011]. The total variability subspace is a low-dimension subspace trained with factor analysis (FA) modeling, where the main variabilities describing the data are kept. Thereby, *iVectors* are a reduced set of measures that capture many aspects of the speech, and our hypothesis is that they also contain information about intelligibility. Thus, intelligibility assessments can be based on them. Our methodology is similar to Bocklet et al. [2012], but instead of using *iVectors*, they used GMM-based supervectors. However, the great dimensionality reduction given by *iVectors* allows building much simpler predictors.

iVectors are computed from the acoustic parametrization of the signal. In our work, the acoustic information is represented through the perceptual linear prediction (PLP) features [Hermansky 1990], a speech representation that encodes frequency information inspired by human speech perception. For every utterance, 39 PLP coefficients are computed every 10ms, and our method produces a single 400-dimensional *iVector* representing the whole utterance instead, independently of its duration.

This work was initiated in Martínez et al. [2013], where we built an intelligibility assessment system for dysarthric speakers based on *iVectors*. We obtained correlations of above 0.90 between the intelligibility ratings given as ground truth and those predicted automatically by our system. The work was done using the Universal Access Speech (UAspeech) database [Kim et al. 2008]. In the present article, this research is extended with new experiments, keeping the same system architecture and configuration. Mainly, and unlike that work, we also analyze the case where we do not have available recordings of the application user for training. In this new scenario, we will see that one big problem of current databases, such as UAspeech, is that each speaker is associated to a single intelligibility rating, and when one speaker is removed from the training dataset, the system has difficulties in learning information about his or her associated intelligibility. To overcome this problem, a classification system is presented, where the intelligibility scale is divided into intervals, with each interval representing a class. The objective is to assign the speaker intelligibility to one of those classes. Paja and Falk [2012] already showed promising results in classification of spastic dysarthric speech of the UAspeech database into two levels of intelligibility: mid-low and mid-high.

Additionally, the system is trained to predict the *Accuracy* given by a speech recognizer when it is used by dysarthric speakers. We realized that the same architecture designed to assess intelligibility could be directly used to predict the performance of a speech recognizer, and that *iVectors* would also capture the information needed to make such predictions. ASR has a high potential for being used by clinicians as an assistive technology for dysarthric speakers, and if we were able to obtain a confidence measure of the recognizer, uptake rates would increase and health-related costs would diminish [Mengistu et al. 2011]. On many occasions, people with dysarthria have limited range of movements, and it can be tough for them to press the keys of a keyboard or move the mouse to use a computer. ASR is an ideal human-computer interaction solution to overcome these difficulties. This problem was tackled in Mengistu et al. [2011] for spastic dysarthric speakers, and they found good correlations between ASR performance and the predictions made by their automatic system using as input linear prediction coefficient (LPC) kurtosis and skewness, LPC residual kurtosis, and F0-range.

One of the main problems of working with dysarthric speech is the scarcity of data within the available databases [Green et al. 2003]. Data recording requires several repetitions of words involving difficult movements of the speech articulators, which can be exhausting to speakers with some dysarthric conditions. In this article, we work with the UAspeech database, where different types of recordings belonging to 15 dysarthric speakers with different degrees of intelligibility are available. Given the limited number of speakers, the experiments conducted in previous studies on this database [Falk et al. 2011, 2012; Christensen et al. 2012; Paja and Falk 2012; Martínez et al. 2013] used data from the test speaker during training (naturally, data not seen in training). This would correspond to a scenario where the users of the assistive technology application were known in advance, and therefore data of the final users could be precollected to build the system. However, although this is common in real life, it is not always the case; there are occasions in which we do not know who will use the system. This opens a window to a more general situation in a clinical environment, where we would desire one single application useful to everybody. In this work, we give a step forward and compare both situations.

The article is organized as follows. In Section 2, the databases used for the experiments are presented, and in Section 3, the system architecture is detailed. The evaluation methods are described in Section 4, and the experiments on intelligibility assessment are shown in Section 5. Section 6 presents the experiments on the ASR *Accuracy* rate assessment, both with regression and classification results. In Section 7, the goodness and efficiency of *iVectors* are analyzed for an intelligibility assessment application. The conclusions of the work are drawn and discussed in Section 8.

Table I. UAspeech Speaker Information

No.	Speaker Label	Age	Speech Intelligibility (%)	Dysarthria Diagnosis
1	M04	>18	Very low (2)	Spastic
2	F03	51	Very low (6)	Spastic
3	M12	19	Very low (7.4)	Mixed
4	M01	>18	Very low (15)	Spastic
5	M07	58	Low (28)	Spastic
6	F02	30	Low (29)	Spastic
7	M16	—	Low (43)	Spastic
8	M05	21	Mid (58)	Spastic
9	F04	18	Mid (62)	Athetoid
10	M11	48	Mid (62)	Athetoid
11	M09	18	High (86)	Spastic
12	M14	40	High (90.4)	Spastic
13	M10	21	High (93)	Mixed
14	M08	28	High (93)	Spastic
15	F05	22	High (95)	Spastic

Note: In the first and second columns, we have the speaker identification number and label. In the third column, we have the speaker's age. In the fourth column, we have the ground truth speech intelligibility ratings given in the UAspeech database. In the fifth column, we have the type of dysarthria of each speaker. We can find spastic (voice perceived as strained, harsh, raspy, slow, showing consonant distortion and hypernasality), Athetoid (unstressed and monotone voice, inappropriate voice stoppage or release, variable speech rate), and mixed (voice shows a combination of effects of previous types).

2. AUDIO MATERIAL

Two databases were used in the training process: UAspeech [Kim et al. 2008] and Wall Street Journal Database 1 (WSJ1) [Paul and Baker 1991]. The first contains dysarthric speech, and the second was used for training maximum likelihood (ML) models with a large number of parameters that require large amounts of data, and it does not contain dysarthric speech. The sampling rate was fixed at 16kHz in both. Next we describe the databases and explain how they were used in this study.

2.1. Universal Access Speech Database

This is a dysarthric speech database recorded from 19 speakers with cerebral palsy. We had data available from 15 speakers. Data were recorded with an eight-microphone array at 48kHz and one digital video camera. For each speaker, 765 words were recorded in three blocks of 255. Of the recorded words, 155 are common to the three blocks (from now on, we will refer to these as the *common* subset) and 100 are uncommon words that differed across them (from now on, we will refer to these as the *uncommon* subset). The 155-word blocks included 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 orthographic transcriptions of each word. The percentages of correct responses for each speaker obtained by the five listeners were averaged to calculate the speaker's intelligibility. In Table I, a summary of each speaker in the database with his or her intelligibility can be seen. For more information about the database, please refer to Kim et al. [2008].

For our experiments, only microphone 6 was used, and two subsets were created: train and test. For testing, we reserved the *uncommon* subset (300 words per speaker), and for training, the rest (465 words per speaker). This configuration, proposed in Falk et al. [2012], permitted us to make fair experiments because the tested words were never seen during training.

2.2. Wall Street Journal 1

WSJ1 is a general-purpose English, large vocabulary, natural language, high-perplexity corpus containing a substantial quantity of speech data (77,800 training utterances totaling about 73 hours of speech). It includes read speech and spontaneous dictation by journalists. The database also contains development and testing datasets in a “Hub and Spoke” paradigm to probe specific areas of interest. Each of them contains 7,500 waveforms—about 11 hours of speech. Data were collected using two microphones at a sampling rate of 16kHz. For more information, please consult Paul and Baker [1991].

This database was selected because it contains a large amount of speech in American English, like UAspeech, so we could train our ML models described in the next section—the Gaussian mixture model (GMM) and FA front-end—more reliably than using only UAspeech. In addition, both databases mostly contain read speech (the words in UAspeech are read from prompts). Only the clean speech of WSJ1 (from its training, development, and testing parts) was used, which totals 73.84 hours.

3. SYSTEM ARCHITECTURE

The architecture presented next can easily be adapted to different applications. We investigated two of great interest to the field of assistive technologies. In the first one, the goal was to assess the intelligibility of dysarthric speakers after they uttered a set of words. In the second one, the goal was to predict the *Accuracy* that a speech recognizer would obtain for those speakers after they uttered those same words. In our experiments, predictions were made over the *uncommon* subset of the UAspeech corpus. These words (e.g., “naturalization,” “moonshine,” “exploit”) were selected from children’s novels digitized by Project Gutenberg, using a greedy algorithm that maximized token counts of infrequent biphones. Thus, they are expected to generalize well and be useful to provide significant metrics of the speakers. In the application training process, they were never seen.

The major novelty of our proposal is that in both cases, the predictions are made from *iVectors*, a compressed acoustic representation containing different aspects of speech. From the viewpoint of a practitioner using assistive technology applications, the most interesting characteristic of *iVectors* is that they capture the intelligibility information of the utterance in a reduced set of measures. From the viewpoint of a researcher building assistive technology applications, their most interesting characteristic is that they are of fixed length and low dimension, and a whole utterance can be represented by a single *iVector* independently of its duration. In our opinion, the most relevant feature of our scheme was that it performed an efficient compression of the acoustic parameters extracted from the speech while keeping the most important information needed to make assessments. The system architecture is depicted in Figure 1. First, PLP coefficients and energy, with their first and second derivatives, were extracted from the speech. Then, a universal background model (UBM) with 1,024 components was trained on WSJ1 and used to compute sufficient statistics of each utterance. Next, the *iVector* extractor was trained with the sufficient statistics calculated for WSJ1. Each *iVector* contained 400 speech measures. Finally, the *iVectors* obtained for the UAspeech database were used for training and evaluating the predictor. In the next sections, every component of the system is explained in detail.

3.1. Acoustic Features

Each audio file was parameterized into 12 PLP features plus energy, with derivatives and accelerations, to obtain a 39-dimension vector every 10ms, in 25ms-length windows. These features use three concepts of the psychophysics of hearing: the critical band spectral resolution, the equal-loudness curve, and the intensity-loudness power

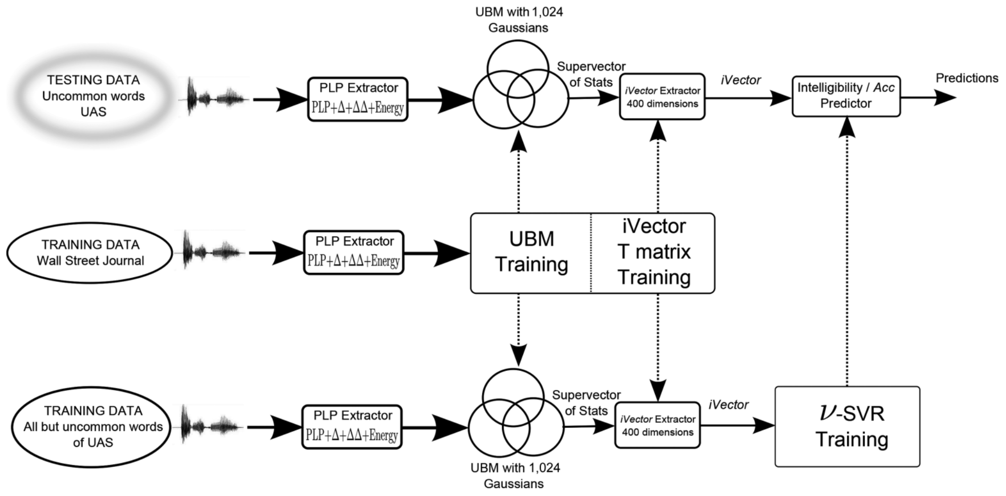

Fig. 1. System architecture. PLPs and supervectors of WSJ1 were used to train the UBM and the *iVector* extractor, respectively. *iVectors* of the training dataset of UAspeech were used to train the predictor. *iVectors* from the *uncommon* subset of the UAspeech database were used for evaluating the system. This subset included 300 words per dysarthric speaker, and each word was represented by one *iVector*.

law [Hermansky 1990; Moore 2003]. Previous investigations showed that there is information about intelligibility in the short-term spectral content [Hosom et al. 2003], but there was no a priori theoretical reason why PLP should work better than other commonly used features, such as mel-frequency cepstral coefficients (MFCCs). However, there are works on speech intelligibility with pathological voices where PLPs offered some advantages over MFCCs [Bocklet et al. 2009]. A reason could be that PLPs follow the peaks of the spectrum better, thanks to the linear prediction (LP) analysis they perform, what is known as the "peak-hugging" property of LP. Then, a better model of the vocal tract transfer function is obtained [Makhoul 1975].

3.2. Gaussian Mixture Model and Sufficient Statistics

A GMM [Reynolds and Rose 1995] is a multimodal distribution typically used in speech processing, where a fixed number of Gaussian components is combined to create a distribution that would be difficult to parameterize with a single function. A GMM was trained on WSJ1 by running 20 iterations of the expectation-maximization (EM) algorithm [Dempster et al. 1977]. This model was our UBM. Once the UBM was trained, zeroth (N)- and first (F)-order Baum-Welch statistics were obtained for each utterance as follows:

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

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

where L is the number of frames in a given file, and $\mathrm{P}(k|x_t, \Omega)$ is the posterior probability of mixture component k generating the PLP vector x_t, for a model with parameters Ω, and K components. A simple intuition behind these vectors is that the zeroth-order statistics count of how many PLPs were generated by each Gaussian component, and the first-order statistics indicate the mean value of the PLPs that belong to each

Gaussian component. The count in zeroth-order statistics is soft in the sense that $P(k|x_t, \Omega)$ is not 1 or 0, but is allowed to be any fractional value in between. This means that each vector is not generated only by a single Gaussian component, but all components are partly responsible for its generation. Supervectors of statistics were subsequently built by concatenating the statistics of each Gaussian component.

3.3. Factor Analysis Front-End: *iVector* Extractor

FA arises as a method to compensate a GMM model for different effects that introduce uncertainty in the speech signal, such as noise, channel, speaker, or the intelligibility of the spoken utterance. In all cases, the underlying model generating the data will be different. Nonetheless, there exists a background model common to all realizations—the UBM—but for each particular case, we have to deviate from the UBM to better match the model generating the current speech data. In other words, we have to add something to the UBM and obtain a new model that better fits every particular speech realization. Those deviations from the UBM contain the peculiarities of the speech signal that we want to retain. They contain the specific information of that utterance about the speaker, the channel, and the intelligibility with which it was uttered. How we move apart from the UBM is (to some extent) what an *iVector* measures. Later, depending on how we group *iVectors*, we can model one aspect of speech or another. This is the reason *iVectors* have also been successful in many areas, such as speaker identification [Dehak et al. 2011] or language recognition [Martínez et al. 2011].

More formally, the main assumption of our FA model was that every utterance s in the database was generated by a different GMM, with mean supervector $m(s)$ modeled as

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

where m_0 is the mean supervector of the UBM, i is a latent variable that has an a priori standard normal distribution $\mathcal{N}(0, 1)$, and T is a JxD matrix that translates *iVectors* from their low-dimension total variability space to the high-dimension space where the model $m(s)$ lies, with D being the *iVector* dimension and $J = 39$xK being the dimension of the supervectors. The *iVector* of utterance s was calculated as the expectation of the posterior distribution of i given the sufficient statistics of the utterance [Dehak et al. 2011]. The main difference between the common formulation of FA [Bishop 2006] and ours is that in the former, the latent variable changes for every frame, whereas in our case, the latent variable is common to the whole utterance [Kenny et al. 2007]. In this way, we obtained a single vector per utterance, the *iVector*, that describes the whole utterance, and not a vector for each frame within the utterance compensating for the variability observed in that frame. The *iVector* contains information about the variability observed in that utterance with respect to the entire database, and our hypothesis was that intelligibility and information correlated to the accuracy of an ASR system are partly responsible for such variability.

The training of T was done with the EM algorithm, alternating an ML step with a minimum divergence step (MD), and the WSJ1 training data were used for this purpose. By using only nondysarthric speech in the *iVector* extraction training, we expected that high intelligible speakers would produce a small shift from the UBM, and low intelligible speakers would produce a large shift from the UBM, which would help to have more discriminative *iVectors*.

3.4. Predictor

The predictor was the block in charge of making assessments from *iVectors*. It can be seen as a block that transforms the *iVector* associated with a given utterance into an intelligibility rating or an *Accuracy* prediction. Note that the labels used in training were the key information to make our system work as an intelligibility assessment system or as an *Accuracy* predictor. Thus, the predictor was the block responsible for extracting the required intelligibility or *Accuracy* information from *iVectors*. Specifically, we used a ν-support vector regression ($\nu - SVR$) predictor [Chang and Lin 2002; Smola and Schölkopf 2004]. The basic idea of SVR is that only a subset of vectors that are not farther than a given margin from the regression curve are used for training. The approximating function is

$$f(\mathbf{x}) = \sum_{i=1}^{N} \hat{\alpha}_{\mathbf{i}} \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) + b, \tag{4}$$

where $\mathbf{x}$ is the target feature, $\hat{\alpha}_{\mathbf{i}}$ is a scalar, b is a bias, N is the number of files in the training dataset, and ϕ is the kernel, in our case a radial basis function, which allows modeling of nonlinearities. The software LIBSVM was used to train the SVR predictor and run tests with it [Chang and Lin 2011]. We set the parameters $C = 1$ and $\nu = 1$. In general, the system was stable in the range of values of $C = 0.5 - 10$ and $\nu = 0.1 - 1$, and the results were not very different within these intervals. Mathematical details about the training of the regressor can be seen in the Appendix.

For training the predictor, we used the *common* subset of the UAspeech database, which includes 465 words per speaker distributed in three blocks of 155 common words. For evaluating the applications, we used the *uncommon* subset of the UAspeech database, which includes 300 different words per speaker. For each word, one *iVector* was extracted and passed directly to the regressor, either for training or testing.

4. APPLICATION EVALUATION METHODS

At the time of building an assistive technology application, it is common that we know in advance the patients who will use the system. This is an ideal scenario because we can use precollected data of those speakers to train our system. However, this is not always the case, and in other situations we do not know who the users of our application will be. Nonetheless, the system should guarantee a high performance in such cases as well. Unfortunately, the performance is not the same. As we will see later in the results, there is a significant difference between having and not having available data of the people who will be evaluated by our system in the training dataset. Traditionally, this second situation has not been studied in the literature due to the scarcity of data and small size of the available databases, and in our experiments we reported results for the two cases.

These two scenarios required two different training strategies. In the case where we included data of the application user in the training dataset, a single predictor was needed, trained on all the dysarthric speakers. In the case where we did not include data of the application user in the training dataset, 15 predictors were built on a leave-one-out strategy, with data of the rest of the dysarthric speakers.

Initially, the intelligibility assessment and *Accuracy* prediction applications were addressed as a regression problem, in which the objective was to predict the exact value of the intelligibility rating or of the *Accuracy* obtained by the speech recognizer given as ground truth. The results obtained by this approach are very informative, since the intelligibility and *Accuracy* scales are continuous, and any value between 0 (very low intelligibility or *Accuracy*) and 100 (very high intelligibility or *Accuracy*) is possible. However, there are different issues that can make continuous ratings misleading. First, a single intelligibility rating per speaker is not a completely fair choice, because the same speaker can utter different phrases with different levels of intelligibility. Second, intelligibility is a subjective measure, and a fixed intelligibility rating cannot be considered as a fixed gold standard, because the same utterance can have different levels of intelligibility for different people. Third, large errors can cause great confusion to the clinician and make him or her believe that the result is much better or much worse than it really is.

To overcome those troubles (at least partially), both tasks were also formulated as a classification problem. An application that classifies intelligibility would just say if the utterance had *very low*, *low*, *mid*, or *high* intelligibility, if four classes were possible. Even simpler, it could just classify utterances into *low* or *high* intelligibility, if only two classes were possible. The same for an *Accuracy* classifier. For the experiments, the speakers were grouped into four or two classes according to their intelligibility or *Accuracy*, by splitting the interval [0,1] into equal parts. Then, the classification was made over the regression results by setting thresholds of 0.25, 0.5, and 0.75 for the four-class problem and 0.5 for the two-class problem. Thus, if we obtained a regression value of 0.60, it belonged to class *mid* in the first case and to class *high* in the second case. Note that in this example, if the true rating were 0.70, it would not have counted as an error in any of the two cases, whereas in the regression approach, the error would not be 0. The classification task was only conducted for the case where we do not have data of the application user to train the system.

The intelligibility and *Accuracy* assessments obtained with the regression system were measured in terms of the following:

—Pearson correlation (r) is a metric that measures the linearity of the relationship between two variables, with 1 meaning perfectly linear, 0 no linear relation, and –1 inverse linear relation. Given that our data were described with parametric models, and that we pursued a linear relationship between rated and predicted intelligibility, this type of correlation was appropriate for our problem. Its mathematical definition can be found in Onwuegbuzie et al. [2007].

—Root mean square error (RMSE) is a measure of the real difference between the predicted and rated values. The smaller this quantity, the closer our predictions are to the subjective ratings. Its mathematical description can be found in Armstrong and Collopy [1992].

—Error rate at 12.5% (error_rate$_{12.5\%}$) is a metric introduced in Martínez et al. [2013] to overcome the subjectivity of intelligibility ratings made by the speech therapists. It shows the percentage of utterances with a prediction error higher or lower than 0.125. The choice of 12.5% is selected to cover intervals including a 25% of the continuous rating scale, the same margin that a hard four-class classification covers. The difference with the four-class classification is that the intervals are not fixed and depend on the target value. It is defined as

$$error\_rate_{12.5\%} = \frac{C^{+} + C^{-}}{N}, \tag{5}$$

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

Classification was measured in terms of weighted average precision and weighted average recall. Precision measures the ratio of true positive outcomes and the sum of all outcomes classified as positive (positive means classified as the class under consideration)-that is, among all outcomes classified as a given class, what percentage really belongs to that class. Meanwhile, recall refers to the ratio of true positive outcomes and all outcomes of that class. This measures the percentage of outcomes of a class correctly classified. These two metrics were measured for each class individually. To obtain a global measure for the system, we averaged the result of all classes, weighting by the number of outcomes of each class. More information about these metrics can be found in Sokolova and Lapalme [2009].

Table II. Pearson Correlation (*r*), RMSE, and Error_Rate$_{12.5\%}$ for the Intelligibility Assessment System When We Had User Data Available in the Training Dataset (Middle Column) and When We Did Not Have User Data Available in the Training Dataset (Right Column)

	User Data in Train	User Data Not in Train
r	0.91	0.74
RMSE	0.14	0.23
Error_rate$_{12.5\%}$	0.33	0.61

All preceding metrics were measured on a per-word (or per-utterance) basis. In other words, Pearson correlation was computed over the intelligibility ratings of all evaluated words; in RMSE and error_rate$_{12.5\%}$, we measured the error of the prediction of each word with regard to its true label; and in classification, we counted the times that each word was correctly assigned to its class. Finally, a single metric for the whole system was obtained by averaging the results of all words. Additionally, we obtained averaged results of the predictions for each speaker, as the ultimate goal of our applications was to obtain intelligibility and *Accuracy* assessments for each evaluated user.

5. EXPERIMENTS ON INTELLIGIBILITY ASSESSMENT

A computer application to automatically obtain intelligibility measures would bring several benefits to clinicians, such as objectivity and replicability of results. It would ensure that speech therapists from different places apply the same criteria to evaluate intelligibility. In addition, clinicians can get used to the speech of their patients and become more familiar with their manner of pronunciation, causing them to provide higher ratings over time even when the speech has not changed. A computer application would also avoid this problem. Intelligibility assessment is an important part in the monitoring of a patient progress, and computers can contribute to perform this task always with the same criteria and with no human subjectivity. In our experiments, we compared an application that was trained with a dataset including prerecorded speech of the evaluated user, with an application that was trained with a dataset without prior information about the user.

5.1. Intelligibility Assessment by Regression

Regression results for intelligibility assessment clearly showed that it was very helpful to count with data of the application user at the time of training the system. The results of this situation are reported in the middle column of Table II; in the right column, we can see the results when data of the application user were not included in training. As we can see, the reduction in correlation and increase of RMSE and error_rate$_{12.5\%}$ were high. Two possible causes were responsible of this behavior, and both arose because of the limitation of the UAspeech database. First, the system was not able to predict intelligibility ratings not seen in training with the same accuracy. Given that we only had 13 different intelligibility labels in the training dataset, if we removed one, the system was not able to interpolate with the rest properly. Second, in the case where the application user was included in the training dataset, the system was learning not only intelligibility information but also the speaker identity. We had very few speakers, and every speaker was uniquely associated to a single label; therefore, each label identified uniquely to that speaker (except for speakers F04 and M11, and speakers M10 and M08, who shared intelligibility rating).

To analyze the results of each speaker individually, the mean and standard deviation of each speaker for the case where we used data of that speaker in the training dataset are plotted (Figure 2) versus the case where we did not use data of that speaker for

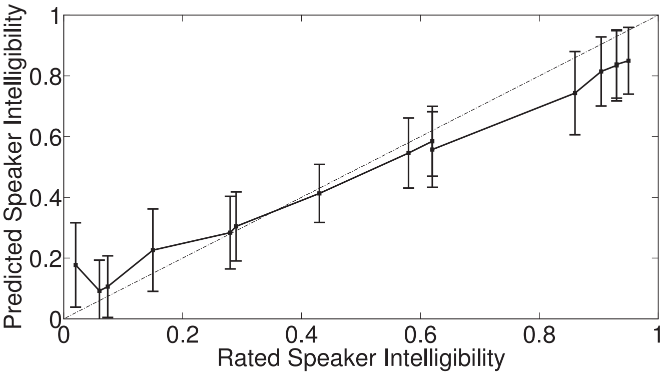

Fig. 2. Mean and standard deviation of intelligibility predictions for each speaker when user data were included in the training dataset (straight) and $x = y$ line (dash-dot).

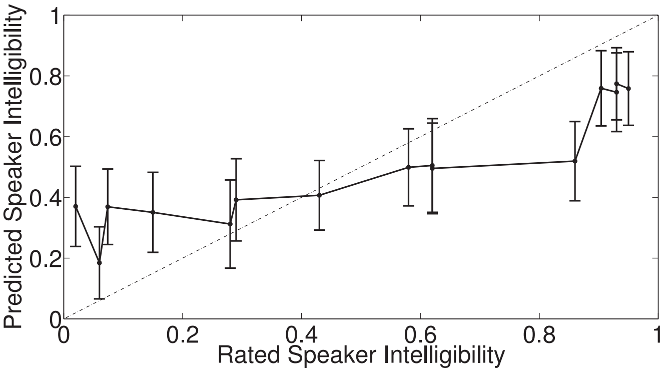

Fig. 3. Mean and standard deviation of intelligibility predictions for each speaker when user data were not included in training dataset (straight) and $x = y$ line (dash-dot).

training (Figure 3). We clearly see that the second curve deviated further from the $x = y$ line (dash-dot curve). This was more dramatic for speakers with very low intelligibility. One could think that this was due to having a UBM and an *iVector* extractor trained without any dysarthric speaker. Then, the less intelligible speakers should have had the least accurate predictions. However, the fact that mid intelligible speakers were better modeled than high intelligible speakers contradicted this hypothesis. In the future, we would like to try a middle solution between training the UBM only with WSJ1 and only with UAspeech, such as a maximum a posteriori (MAP) adaptation [Gauvain and Lee 1994] from the WSJ1 UBM using dysarthric speakers.

We must also take into account the data scarcity issue, and when we did not include the evaluated speaker in the training dataset, we lost a nonnegligible part of the training data. The consequences of this were even more important for the least and most intelligible speakers, for whom the system could not interpolate with any other speakers to learn their associated intelligibility. Therefore, including the test speaker in training can be thought as an ideal scenario, where all speakers were perfectly represented in the training dataset. This yielded optimal results, as observed.

Despite all of these problems, the correlation obtained for the case where we did not include data of the application user in the training dataset was of about 0.74, which can still be interpreted as a high correlation. The comparison in Table II could be considered to be unfair, because in the case where we had data of the evaluated

Table III. Four-Class Intelligibility Assessment Classification Confusion Matrix in Percentage of Words

Label (↓)\Decision (→)	Very Low	Low	Mid	High
Very Low	31.76	57.69	10.19	0.37
Low	19.55	61.65	18.67	0.13
Mid	4.12	46.88	44.38	4.62
High	0.23	10.50	44.86	44.41

Table IV. Two-Class Intelligibility Assessment Classification Confusion Matrix in Percentage of Words

Label (↓)\Decision (→)	Low	High
Low	85.94	14.06
High	25.89	74.11

speaker in the training dataset, we trained the predictor with more data compared to the case where there was no information about the evaluated speaker in the training dataset. To investigate this, we carried out a control experiment including information of the application user, but where the amount of data was reduced (we removed the same amount of data per speaker), to match that available for the case with no data of the application user in the training part. The results showed no difference, and $r = 0.913$ and $RMSE = 0.140$ were obtained in this scenario.

One interesting point would be to see if there were some words better predicted than others, and to analyze if the best predictions came from particular word patterns. We measured the difference between the rated and automatically predicted intelligibility for each of the 300 tested words per speaker, and most of them were in the range of 0.14 to 0.24. The worst predicted words, with a mean difference computed over all speakers higher than 0.25, were *behavior*, *employment*, *scissors*, *aloft*, *booth*, *buffoon*, *fishing*, *swoon*, and *ahead*. The best predicted words, with a mean difference over all speakers under 0.13, were *Pennsylvania*, *advantageous*, *bloodshed*, and *designate*. We did not find any phonetic cue indicating that some patterns were better rated than others. However, it seems that shorter words were more difficult to predict, although there was no strong evidence of this.

5.2. Intelligibility Assessment by Classification

The problems associated with regression caused by having each speaker associated with a unique intelligibility rating should be alleviated in a classification problem, because several speakers with different ratings are grouped into the same class. In addition, the interpolation of extreme intelligibility ratings should not be as crucial, and it will be sufficient if the system learns that the *iVector* associated to an utterance is close to others of the same class. The classification problem was conducted only for the case where there was no data of the application user in the training dataset.

The results of the four-class classification problem are given in Table III in the form of a confusion matrix. Encouragingly, the confusions primarily were made with neighboring classes. This fact assures that there was no overtraining of any class. The problem was difficult though, especially for the *very low* class, for which only 31.76% of the files were correctly classified, and almost 60% of the files were confused with the *low* class. The average weighted precision was 0.60, and the average weighted recall was 0.44.

Given that there was still large confusions with neighbor classes, a simpler and more reliable solution for real applications would be to just discriminate between *high* and *low* intelligibility. As we can see in Table IV, the results of this scenario were much better, but also note that the classes were twice as wide as for the four-class problem,

Table V. *Accuracy* Labels and Assessments Per Speaker

No.	Speaker Label	Speech Intelligibility (%)	*Accuracy* (%) in Christensen et al. [2012]	Mean *Accuracy* (%) Assessment User Data in Train	Mean *Accuracy* (%) Assessment User Data Not in Train
1	M04	Very low (2)	8.30	25.61	49.41
2	F03	Very low (6)	23.00	23.37	30.53
3	M12	Very low (7.4)	11.70	15.18	44.81
4	M01	Very low (15)	29.80	34.08	42.15
5	M07	Low (28)	66.90	55.12	35.43
6	F02	Low (29)	36.90	37.99	43.79
7	M16	Low (43)	49.30	49.61	48.82
8	M05	Mid (58)	53.40	53.12	51.95
9	F04	Mid (62)	65.60	61.71	53.47
10	M11	Mid (62)	53.00	52.07	54.30
11	M09	High (86)	81.50	72.39	54.65
12	M14	High (90.4)	74.90	72.39	71.59
13	M10	High (93)	86.20	78.19	69.50
14	M08	High (93)	81.80	76.32	70.95
15	F05	High (95)	89.60	80.77	68.97
Total			54.10	52.69	52.69

Note: In the first two columns, we have the speaker identification number and label. In the third column, we have the speech intelligibility ratings given in the UAspeech database. In the fourth column, we have the *Accuracy* labels obtained in Christensen et al. [2012] (our ground truth for the *Accuracy* prediction task). In the fifth column, we have the *Accuracy* predictions obtained by our system when there was user data available in the training dataset. In the sixth column, we have the *Accuracy* predictions obtained by our system when there was not user data available in the training dataset. In the last row, we have the averages obtained for the fourth, fifth, and sixth columns.

and hence the information given by the system was not as precise as the one given by the four-class classification system. Both weighted precision and recall were 0.80.

6. EXPERIMENTS ON AUTOMATIC SPEECH RECOGNITION WORD ACCURACY RATE ASSESSMENT

The interest of this application lies in obtaining confidence measures of ASR systems that guarantee successful usage. ASR has the potential to be a very important human-computer interaction mechanism for people with limited range of movements, as are many people affected by dysarthria. Again, we compared the cases where data of the application user were and were not available in advance for training. The only change with respect to the intelligibility assessment experiment was the use of different labels to train and test the system. Instead of the intelligibility ratings given by the UAspeech database, our ground truth labels were the *Accuracies* obtained by the reference speech recognizer, which was the *mapSI2* system presented in Christensen et al. [2012]. That was the best-performing system among 11 presented in that paper evaluating the same dysarthric speakers as we did. In that work, the authors used data of the evaluated speakers to build the recognizer and made a MAP adaptation to the final user. The ground truth *Accuracies* for the 15 speakers of the UAspeech database are reflected in Table V.

6.1. Word Accuracy Rate Assessment by Regression

In general, this task was more complicated than intelligibility assessment, especially for the *very low* intelligible speakers, for whom better *Accuracies* than the real ones were obtained. Observe in Table V that there were only three speakers with *Accuracy*

Table VI. Pearson Correlation (*r*), RMSE, and Error_Rate$_{12.5\%}$ for the *Accuracy* Prediction System When We Had User Data Available in the Training Dataset (Middle Column) and When We Did Not Have User Data Available in the Training Dataset (Right Column)

	User Data in Train	User Data Not in Train
r	0.89	0.55
RMSE	0.12	0.22
Error_rate$_{12.5\%}$	0.26	0.56

Table VII. Four-Class *Accuracy* Classification Confusion Matrix in Percentage of Words

Label (↓)\Decision (→)	Very Low	Low	Mid	High
Very Low	9.95	64.18	25.62	0.25
Low	3.35	64.68	31.97	0.00
Mid	3.89	37.05	49.03	0.10
High	0.00	8.91	68.82	22.27

Table VIII. Two-Class *Accuracy* Classification Confusion Matrix in Percentage of Words

Label (↓)\Decision (→)	Low	High
Low	71.07	28.93
High	26.81	73.19

below 25% (M04, F03, and M12); therefore, it was very hard for the regressor to learn representative patterns of the lowest *Accuracies*. Note that in this case, the labels were the result of the filtering process committed by the speech recognizer, which was not error free. This means that the labels might not be completely accurate. Despite this problem, when data of the application user were included in the training dataset, a correlation of about 0.90 was obtained, as can be seen in the middle column of Table VI. When no data of the application user were included in the training dataset, the drop in the correlation was dramatic, and the RMSE and error_rate$_{12.5\%}$ increased significantly, as observed in the right column of the same table. However, RMSE and error_rate$_{12.5\%}$ were smaller than in the intelligibility assessment task. This result could be misleading, because although the *very low* and *very high* predictions were less accurate, we had fewer speakers labeled with these *Accuracies*; this caused a smaller value of RMSE and error_rate$_{12.5\%}$. Consequently, the behavior of the intelligibility assessment system was preferred, because the system was not biased to predict some intervals more or less likely than others. In Table V, the results of the *Accuracy* predictions for each speaker are shown for the cases where the user data were and were not present in the training dataset.

6.2. Word Accuracy Rate Assessment by Classification

The same problem as for the regression approach was observed in *Accuracy* classification. The *very low* class was not modeled well, resulting in only 10% of the utterances well classified in the four-class classification problem. In general, the system was biased to predict the *low-mid* interval, as can be seen by the confusions in Table VII. For example, for the *very low* class, there were more words classified as *mid* than as *very low*. An average weighted precision of 0.45 and a weighted recall of 0.37 were obtained. Remember that the classification problem was conducted only for the case where data of the application user was not included in the training dataset.

A more reliable solution was the system that classifies just between *low* and *high*. In Table VIII, we have the confusion matrix of this two-class classification problem. In this

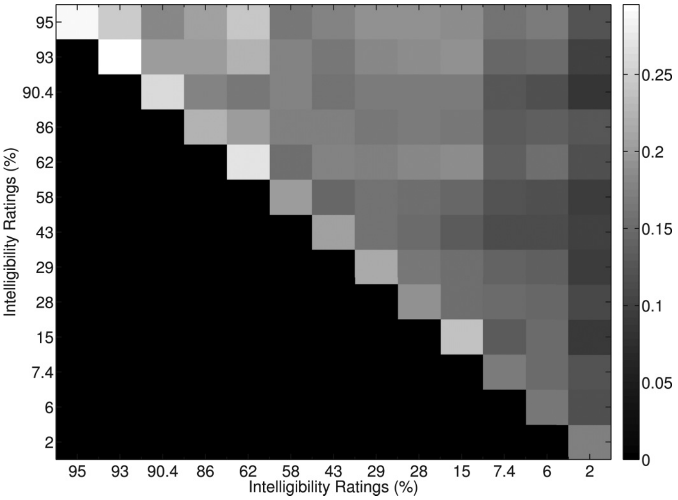

Fig. 4. Matrix of average CDS calculated from *iVectors* among all possible intelligibility pairs. Each cell of the matrix shows the average CDS of the intelligibility rating pair indicated by the corresponding row and column. The whiter the cell, the closer the *iVectors*.

case, both weighted precision and recall were 0.72. These results confirm that *Accuracy* prediction was more challenging than intelligibility assessment, in our opinion because the labels were noisier due to the filtering process that the speech recognizer made.

7. ANALYSIS OF *iVECTORS*

It is interesting to observe the potential of *iVectors* and to know if the proposed method is better than other techniques previously used in the literature. In this section, we studied the goodness of *iVectors* and analyzed why we obtained the reported results. Then, we compared the results with two systems without *iVectors*. In the first, we extracted the mean of the PLPs of each utterance and assessed intelligibility directly with the resulting set of coefficients. In the second, we used the supervectors calculated with the UBM directly to make the assessments, similarly to Bocklet et al. [2012].

7.1. Goodness of *iVectors*

In this section, we discuss our analysis of the goodness of *iVectors* to do intelligibility assessments, in which case we studied the similarity between *iVectors* extracted for all intelligibility ratings. The basic idea was to see if *iVectors* of the same rating were similar and how similar they were, and if *iVectors* of different ratings differed and how much they differed. The metric used to measure this similarity was the cosine distance scoring[1] (CDS).

First, we computed CDS between all possible pairs of *iVectors* extracted from the UAspeech database (including training and testing datasets). In other words, we measured the similarity between all words included in the database. Then, for every possible intelligibility pair, we averaged all CDSs calculated with all *iVector* pairs corresponding to speakers rated with those two intelligibilities. The result can be seen in Figure 4. We obtained a matrix where the average CDS among *iVectors* belonging to the intelligibility rating i and the intelligibility rating j is shown in row ith and column jth. Therefore, in the main diagonal, the CDS among *iVectors* belonging to the same

[1] For *iVectors* A and B, the CDS is defined as the cosine of the angle Θ between them, $CDS = cos(\Theta) = \frac{A \cdot B}{||A||||B||}$, thus higher CDS means more similar *iVectors*.

Table IX. Pearson Correlation (*r*), RMSE, and Error_Rate$_{12.5\%}$ for the Intelligibility Assessment System Based on PLP Means When We Had User Data Available in the Training Dataset (Middle Column) and When We Did Not Have User Data Available in the Training Dataset (Right Column)

	User Data in Train	User Data Not in Train
r	0.83	0.28
RMSE	0.19	0.34
Error_rate$_{12.5\%}$	0.45	0.77

intelligibility rating is plotted. Note that there were 15 speakers and 13 ratings, because there were two pairs of speakers sharing the same rating (F04 and M11, and M10 and M08). The matrix is upper triangular to avoid replicating the information twice. The lower part would be the upper part transposed.

It can be observed that the similarity between high intelligibility pairs was higher (higher CDS) than that of the low intelligibility pairs—that is, the left upper part of Figure 4 is lighter than the right lower part, which is darker. In addition, *iVectors* belonging to high intelligibility ratings were not similar to those of low intelligibility ratings—that is, the right upper part is dark. As well, it can be seen that the main diagonal decreases progressively from whiter to darker. This means that *iVectors* belonging to high intelligibility ratings were more similar among themselves than *iVectors* belonging to low intelligibility ratings. This is a normal behavior, as very severe dysarthric speakers can produce very different sounds even when they want to say the same word or sentence. This shows that the sound variability of very low intelligible dysarthric speakers was higher than that of more intelligible dysarthric speakers. Therefore, the progressive decrease of the CDS indicated that *iVectors* varied consistently as we passed from high to low intelligibility. Ideally, we would like to have a main diagonal as white as possible, indicating that *iVectors* belonging to the same intelligibility rating are very similar. In conclusion, *iVectors* behaved as expected, and they have the potential to be good features for intelligibility assessment. The results will improve if we are able to obtain *iVectors* more similar when they belong to the same intelligibility rating and less similar when they belong to different intelligibility ratings. The difference between high and low intelligibility speakers arose as a natural consequence of very low intelligible speakers being less consistent in their realizations.

7.2. Intelligibility Assessment with Perceptual Linear Prediction Means

The simplest approach to assess intelligibility that we could think of was to compute the mean of the PLPs of each file and assess intelligibility with the resulting set of coefficients. In this way, every word was represented by the mean of the PLPs, and that mean was the input to the regressor. This allowed us to check if *iVectors* were really keeping the important information while compressing the acoustic parameters. As we can see in Table IX, for the case where user data were available in the training dataset, the results with this method were worse than with the *iVector* system, but they were still good. However, for the case where we did not include data of the application user in the training dataset, the decrease in performance was dramatic. This confirmed that *iVectors* were working.

7.3. Intelligibility Assessment with Supervectors of First-Order Statistics

Another interesting experiment to see if *iVectors* were really effective was to compare the intelligibility assessment *iVector*-based system with a system where, similarly to Bocklet et al. [2012], intelligibility was assessed directly with supervectors of first-order statistics extracted with the UBM, as defined in (2). In this approach, the *iVector*

Table X. Pearson Correlation (r), RMSE, and Error_Rate$_{12.5\%}$ for the Intelligibility Assessment System Based on First-Order Statistic Supervectors When We Had User Data Available in the Training Dataset (Middle Column) and When We Did Not Have User Data Available in the Training Dataset (Right Column)

	User Data in Train	User Data Not in Train
r	0.90	0.71
RMSE	0.15	0.24
Error_rate$_{12.5\%}$	0.35	0.60

extractor block was removed from our scheme, but we did not get the same big compression rates. The results are in Table X. As we can see, *iVectors* allowed increasing the system performance, and even more important, they allowed a big reduction in computational time and simplicity. The dimension of supervectors was very high ($1024 \cdot 39 = 39936$), and the regressor had more problems learning the important information. *iVectors* removed noisy information from supervectors and kept important information, and this produced better results.

One interesting observation is that the system with PLP means as input behaved well when there was user data available in the training dataset, but when there was not, the results dropped dramatically. This might be an indication that when data of the evaluated user was included in training, the system learned speaker information that the PLP means efficiently collected instead of intelligibility information. However, the *iVector* and supervector of statistics systems did not suffer such a dramatic drop. This might suggest that these two approaches really learned intelligibility information.

8. DISCUSSION AND CONCLUSIONS

Two interesting assistive applications for people with dysarthric speech were proposed in this article based on *iVectors*. The first had the goal of making automatic intelligibility assessments of dysarthric speech. In our application, the intelligibility rating of the person was made from a set of words not seen during the training process. The importance of making automatic intelligibility assessment for clinicians monitoring the progress of their patients is huge. It would allow making objective assessments that are easily replicated. Furthermore, practitioners involuntarily get used to the speech of their patients, and such a tool would avoid this problem. Unlike humans, our application is not retrained every time the patient speaks.

The second application was designed to predict the *Accuracy* that a speech recognizer will obtain for dysarthric speakers. Speech recognizers have a great potential to be used by disabled people with a limited range of movements, as is the case for many dysarthric speakers. They can serve as a human-computer interface when other common devices, such as keyboards or mice, cannot be used. If predictions of how the speech recognizer will perform for each user could be known beforehand, health costs and abandoned usage rates would be diminished. As for the intelligibility assessment application, the user only had to utter a set of words to obtain a prediction.

One of the main problems for the extension of voice interfaces for dysarthric speakers is that they cannot use standard applications. Dysarthric speech contains a high degree of variability, and such complicated inputs are very difficult to handle by current ASR systems. Likely, the cheapest solution for this would be to convert the dysarthric speech into a more intelligible one so that current ASR systems could understand it [Hosom et al. 2003; Kain et al. 2007]. However, a technologically easier option is to train specific systems for people with this type of disability. At the present time, some progress has been made in this direction [Hamidi et al. 2010; Christensen et al. 2012]. Nevertheless, we have to bear in mind that the extension of these applications will be limited to the individuals for whom it was trained, and that they require a recording process, which is often hard and very exhausting for people with disabilities. Intelligibility assessment can play an important role in these tasks. It can be used to determine if a dysarthric speaker will be able to use a standard ASR system or if it is better to use a specific application designed with consideration of his or her limitations. In that sense, this functionality is similar to that given by an ASR *Accuracy* prediction application, but at training the latter must decide the type of speech recognizer whose performance is going to be predicted (global or specific for dysarthric speakers).

A very relevant conclusion of our work is that it is helpful to include precollected data of the application user for training our applications. In a clinical environment, it is common to know who will use the application in advance. However, this is an ideal scenario that is not always possible. Unfortunately, when data of the application user is not included for training, the performance of our systems dropped significantly. The implications of these results are of great impact among researchers and practitioners. They show a handicap for the extension of voice interfaces among dysarthric people, and more effort is needed to improve results in such situations.

For the case of intelligibility assessment, the correlation between intelligibility perceptual ratings and the automatic intelligibility assessments made by our application fell from about 0.90, when we had user data available in the training dataset, down to about 0.74, when we did not, whereas the RMSE increased from 0.14 to 0.23. The assessments were worse for the *very low* and *high* intelligible speakers than for the *mid* intelligible ones. One important reason was data scarcity, especially for the speakers with extreme intelligibility ratings, because the regressor had no information of lower or higher intelligibility ratings with which to interpolate.

For the case of *Accuracy* predictions, the correlation between the true scores obtained with the speech recognizer and the automatically predicted *Accuracies* made by our application fell from about 0.89, when we had user data available in the training dataset, down to 0.55, when we did not, whereas the RMSE increased from about 0.12 to 0.22. In this case, worse results were also obtained for the speakers with *very low* and *high* intelligibility, but especially for the *very low*, because there were only three speakers in the database labeled with *very low Accuracy*, and the system did not have enough information to model them properly. Moreover, for this application, the ground truth labels came from the evaluation of a speech recognizer, which was not error free. Hence, it is likely that these labels were not completely accurate.

iVectors were used as a method to compress the acoustic parametrization of the signal. They capture many aspects of the speaker's speech in a reduced set of measures, in our work 400, instead of 39 PLP coefficients extracted every 10ms. Note that with PLPs, we would have 390 parameters in only 100ms of speech. Their ability to capture intelligibility information and the *Accuracy* that a speech recognizer would obtain was shown with the experiments of this work. *iVectors* were extracted with an FA model. The main difference between our FA and a traditional FA was that the variability modeled in the low-dimension subspace is considered on a per-utterance basis instead of a per-frame basis. That means that each audio recording was represented by a single *iVector*. Thus, simple predictors can be used with them, such as ν-SVR.

Regarding the usefulness of *iVectors*, it was shown that they fulfilled the desired conditions to capture intelligibility information. One interesting observation was that *iVectors* belonging to *low* intelligible speakers were more different among them than *iVectors* belonging to *high* intelligible speakers. This is due to the variability in the speech production of severe dysarthric speakers, whose utterances can vary a lot from realization to realization, even if they say the same word. Finally, *iVectors* were compared with other approaches to assess intelligibility. The conclusion was that *iVectors* kept better the important information to make intelligibility assessments and *Accuracy* predictions while performing a more efficient compression. Hence, we will continue our research in this direction.

In the present work, the acoustic information was captured with the PLP coefficients. These features are not especially designed for intelligibility assessment, and we think that the results could be improved by adding other features more specific to this task. However, our results were competitive and comparable to other works in the literature, such as Falk et al. [2012], where $r = 0.94$ and $RMSE = 0.186$ were obtained over only 10 spastic speakers from the UAspeech database, using six features representing atypical vocal source excitation, temporal dynamics, and prosody. In that work, the information of the application user was also included in training.

APPENDIX

In this appendix, we include the dual formulation problem of $\nu - SVR$ [Chang and Lin 2002; Smola and Schölkopf 2004], defined as

$$\begin{gathered} min\frac{1}{2}(\alpha-\alpha^*)^T\mathbf{Q}(\alpha-\alpha^*)+\mathbf{y}^T(\alpha-\alpha^*) \\ \mathbf{e}^T(\alpha-\alpha^*)=0, \mathbf{e}^T(\alpha-\alpha^*)\leq C\nu, \\ 0\leq\alpha_i,\alpha_i^*\leq C/N, i=1\ldots N, \end{gathered} \tag{6}$$

where N is the number of files in the training dataset; α and α^* are Lagrange multipliers; $\mathbf{e}$ is the vector of all ones; $\mathbf{y}$ represents the target values; C is the regularization parameter; ν is a parameter that controls the number of support vectors and training errors, and unlike ϵ-SVR [Smola and Schölkopf 2004], it avoids a direct selection of the interval around the target values where errors do not count; and $\mathbf{Q_{ij}} \equiv \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ is the kernel, with x_i and x_j the training features. A radial basis function is used as kernel

$$\phi(\mathbf{x_i},\mathbf{x}) = e^{-\gamma||\mathbf{x_i}-\mathbf{x}||_2^2}, \tag{7}$$

where $\gamma = 1/D$, with D the feature dimension. Then, the approximating function is

$$f(\mathbf{x}) = \sum_{i=1}^{N}\hat{\alpha}_\mathbf{i}\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)+b, \tag{8}$$

where $\hat{\alpha}_\mathbf{i} = \alpha_\mathbf{i} - \alpha_\mathbf{i}{}^*$, $\mathbf{x}$ is the target feature, and b is a bias defined in the primal problem.

ACKNOWLEDGMENTS

This work was initiated during an internship of David Martínez in the SPandH group of the University of Sheffield.

REFERENCES

J. Scott Armstrong and Fred Collopy. 1992. Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting* 8, 1, 69–80.

Cristopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer Series in Information Science and Statistics. Springer. DOI:http://dx.doi.org/10.1117/1.2819119

Tobias Bocklet, Tino Haderlein, Florian Hönig, Frank Rosanowski, and Elmar Nöth. 2009. Evaluation and assessment of speech intelligibility on pathological voices based upon acoustic speaker models. In *Proceedings of the 3rd Advanced Voice Function Assessment International Workshop*. 89–92.

Tobias Bocklet, Korbinian Riedhammer, Elmar Nöth, Ulrich Eysholdt, and Tino Haderlein. 2012. Automatic intelligibility assessment of speakers after laryngeal cancer by means of acoustic modeling. *Journal of Voice* 26, 3, 390–397. DOI:http://dx.doi.org/10.1016/j.jvoice.2011.04.010

James Carmichael. 2007. *Introducing Objective Acoustic Metrics for the Frenchay Dysarthria Assessment Procedure*. Ph.D. Dissertation. University of Sheffield.

James Carmichael and Phil Green. 2004. Revisiting dysarthria assessment intelligibility metrics. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP'04)*.

Chih-Chung Chang and Chih-Jen Lin. 2002. Training v-support vector regression: Theory and algorithms. *Neural Computation* 14, 8, 1959–1977.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2, 3, 27:1–27:27. http://www.csie.ntu.edu.tw/∼cjlin/libsvm/.

Heidi Christensen, Stuart Cunningham, Charles Fox, Phil Green, and Thomas Hain. 2012. A comparative study of adaptive, automatic recognition of disordered speech. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH'12)*. 1776–1779.

Frederic L. Darley, Arnold Elvin Aronson, and Joe Robert Brown. 1975. *Motor Speech Disorders*. W. B. Saunders.

Marc De Bodt, María Hernández-Díaz Huici, and Paul H. Van De Heyning. 2002. Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of Communication Disorders* 35, 3, 283–292.

Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2011. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 4, 788–798. DOI:http://dx.doi.org/10.1109/TASL.2010.2064307

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39, 1, 1–38.

Philip C. Doyle, Herbert A. Leeper, Ava-Lee Kotler, Nancy Thomas-Stonell, Charlene O'Neill, Marie-Claire Dylke, and Katherine Rolls. 1997. Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility. *Journal of Rehabilitation Research and Development* 34, 3, 309–316.

Sakina S. Drummond. 1993. *Dysarthria Examination Battery*. Communication Skill Builders.

Pam Enderby. 1983. *Frenchay Dysarthria Assessment*. College Hill Press.

Pam Enderby. 2013. Disorders of communication: Dysarthria. In *Handbook of Clinical Neurology* (110 ed.). Elsevier B. V., 273–281.

Tiago H. Falk, Wai-Yip Chan, and Fraser Shein. 2012. Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. *Speech Communication* 54, 5, 622–631. DOI:http://dx.doi.org/10.1016/j.specom.2011.03.007

Tiago H. Falk, Richard Hummel, and Wai-Yip Chan. 2011. Quantifying perturbations in temporal dynamics for automated assessment of spastic dysarthric speech intelligibility. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'11)*. 4480–4483.

Jean-Luc Gauvain and Chin-Hui Lee. 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions os Speech and Audio Processing* 2, 2, 291–298.

Phil Green, James Carmichael, and Athanassios Hatzis. 2003. Automatic speech recognition with sparse training data for dysarthric speakers. In *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH'03)*. 1189–1192.

Foad Hamidi, Melanie Baljko, Nigel Livingston, and Leo Spalteholz. 2010. CanSpeak: A customizable speech interface for people with dysarthric speech. In *Proceedings of the International Conference on Computers Helping People with Special Needs*. 605–612. DOI:http://dx.doi.org/10.1007/978-3-642-14097-6_97

Hynek Hermansky. 1990. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustic Society of America* 87, 4, 1738–1752.

John-Paul Hosom, Alexander B. Kain, Taniya Mishra, Jan P. H. van Santen, Melanie Fried-Oken, and Janice Staehely. 2003. Intelligibility of modifications to dysarthric speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*. I:924–I:927. DOI:http://dx.doi.org/10.1109/ICASSP.2003.1198933

Alexander B. Kain, John-Paul Hosom, Xiaochuan Niu, Jan P. H. van Santen, Melanie Fried-Oken, and Janice Staehely. 2007. Improving the intelligibility of dysarthric speech. *Speech Communication* 49, 9, 743–759. DOI:http://dx.doi.org/10.1016/j.specom.2007.05.001

Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel. 2007. Speaker and session variability in GMM-based speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 4, 1448–1460. DOI:http://dx.doi.org/10.1109/TASL.2007.894527

Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas Huang, Kenneth Watkin, and Simone Frame. 2008. Dysarthric speech database for universal access research. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH'08)*. 1741–1744.

John Makhoul. 1975. Linear prediction: A tutorial review. *Proceedings of the IEEE* 63, 4, 561–580.

David Martínez, Phil Green, and Heidi Christensen. 2013. Dysarthria intelligibility assessment in a factor analysis total variability space. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH'13)*.

David Martínez, Oldrich Plchot, Lukás Burget, Glembek Ondrej, and Pavel Matejka. 2011. Language recognition in iVectors space. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH'11)*. 861–864.

Knife T. Mengistu, Frank Rudzicz, and Tiago H. Falk. 2011. Using acoustic measures to predict automatic speech recognition performance for dysarthric speakers. In *Proceedings of Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA'11)*.

Catherine Middag, Tobias Bocklet, Jean-Pierre Martens, and Elmar Nöth. 2011. Combining phonological and acoustic ASR-free features for pathological speech intelligibility assessment. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH'11)*. 3005–3008.

Catherine Middag, Jean-Pierre Martens, Gwen Van Nuffelen, and Marc De Bodt. 2009. Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Advances in Signal Processing* 2009, Article No. 3. DOI:http://dx.doi.org/10.1155/2009/629030

Brian Moore. 2003. *An Introduction to the Psychology of Hearing* (5th ed.). Academic Press.

Anthony J. Onwuegbuzie, Larry Daniel, and Nancy L. Leech. 2007. Pearson product-moment correlation coefficient. In *Encyclopedia of Measurement and Statistics*. Sage Publications, 750–755.

Milton S. Paja and Tiago H. Falk. 2012. Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH'12)*.

Douglas B. Paul and Janet M. Baker. 1991. The design for the Wall Street Journal–based CSR corpus. In *Proceedings of the Workshop on Speech and Natural Language (HLT'91)*.

Douglas A. Reynolds and Richard C. Rose. 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing* 3, 1, 72–83.

Sandra J. Robertson. 1982. *Dysarthria Profile*. Winslow Press.

Harsh V. Sharma, Mark Hasegawa-Johnson, Jon Gunderson, and Adrienne Perlman. 2009. Universal access: Preliminary experiments in dysarthric speech recognition. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH'09)*. 7–10.

Alex J. Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Computing* 14, 199–222.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing and Management* 45, 4, 427–437. DOI:http://dx.doi.org/10.1016/j.ipm.2009.03.002

Edythe A. Strand. 2004. Dysarthrias: Management. In *The MIT Encyclopedia of Communication Disorders*. MIT Press, Cambridge, MA, 129–132.

Gwen Van Neuffelen, Catherine Middag, Marc De Bodt, and Jean Pierre Martens. 2009. Speech technology-based assessment of phoneme intelligibility in dysarthria. *International Journal of Language and Communication Disorders* 44, 5, 716–730. DOI:http://dx.doi.org/10.1080/13682820802342062

Received April 2014; revised February 2015; accepted March 2015