Perceptual evaluation for automatic anomaly detection in disordered speech: Focus on ambiguous cases

Imed Laaridh, Christine Meunier, Corinne A Fredouille

Perceptual Evaluation for Automatic Anomaly Detection in Disordered Speech : focus on ambiguous cases

Imed Laaridh[a,b], Christine Meunier[b], Corinne Fredouille[a]

[a] *University of Avignon, LIA, France*
[b] *Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France*

Abstract

Perceptual evaluation is still the most common method in clinical practice for diagnosing and following the condition progression of people suffering from dysarthria (or speech disorders more generally). Such evaluations are frequently described as non-trivial, subjective and highly time-consuming (depending on the evaluation level). Most of the time, perceptual assessment is performed individually by clinicians which can be problematic since judgment may vary from one clinician to the other. Clinicians have therefore expressed the need for new objective evaluation tools better adapted to longitudinal studies, the observation of small units and rehabilitation context to monitor patients' progress. We have previously proposed an automatic approach to the anomaly detection at the phone level for dysarthric speech. The system behavior was studied and validated with different corpora and speech styles and shows good results in this specific task. Nonetheless, the lack of annotated French dysarthric speech corpora has limited our ability to analyze some aspects of its behavior, such as severity, more precisely (more anomalies are detected automatically compared with human experts). To overcome this limitation, we proposed an original perceptual evaluation protocol applied to a limited set of decisions made by the automatic system, relating to the presence of anomalies. Particularly, we intended to focus our analyses on some ambiguous cases in order to enrich our knowledge about the

Email addresses: imed.laaridh@alumni.univ-avignon.fr; laaridh.imed@gmail.com (Imed Laaridh), christine.meunier@univ-amu.fr (Christine Meunier), corinne.fredouille@univ-avignon.fr (Corinne Fredouille)

system behavior. This evaluation was carried out by a jury of 29 non-naive individuals. Results confirm the relevance of the system for the anomaly detection, and place it within the most severe juries. Besides interesting information related to the system behavior, the evaluation protocol highlighted main differences between human process and the automatic system: humans have difficulties in focusing on small units and they are influenced by contextual information, while the system only focuses on small units. In a way, both approaches are probably complementary.

Keywords: Dysarthria, speech disorders, automatic speech processing, perceptual evaluation

---

1. Introduction

Dysarthria: clinical context and limitations

Dysarthria is a motor speech disorder that is a consequence of neurological damage located either in the central or in the peripheral nervous system. This may result in disturbances in any of the components involved in speech production, such as respiration, phonation, articulation, and prosody. Consequently, this may be reflected by weakness, spasticity, incoordination, involuntary movements, or abnormal muscle tone [1, 2, 3], depending on the location of the neurological damage.

Dysarthric speech has been studied according to different axes and objectives. The pioneer studies, conducted in [4, 1, 5] and pursued in [3], relied on the assumption of an unequivocal association between targeted neurological damage and a set of perceptual alterations in speech production. In these studies, the perceptual evaluation of speech, based on 38 auditory-perceptual features and conducted on a large population of dysarthric speakers, leads to the well-known Mayo clinic classification of dysarthria. This classification is still used in clinical practice for assessment and diagnosis of speaker dysarthria. Indeed, perceptual evaluation by one, or a set of listeners, is still the most common paradigm used to evaluate the characteristics and severity of impairment in speech pathologies.

Other studies focused on measuring the degree of severity of the dysarthria. Such a measurement could be defined according to the patient's intelligibility, comprehensibility, efficiency and perceptual degrees of severity [6]. Orthographic transcription of speech samples (sentences, words, pseudo-words) is

also considered a standard method of assessing intelligibility [7] of pathological speech. Such methods are highly variable considering the different granularities (phoneme, syllable or word, sentence) and speech production tasks (read speech, isolated words, pseudo-words or selection of the pronounced word from a closed list of possible productions, etc.) that could be used.
The clinical evaluation of patients is based on several batteries of tests in which the production of dysarthric speakers is rated perceptually by clinicians. According to the different existing evaluation scales, speech evaluation can be quantitative or qualitative. Quantitative scales assess whether or not one of the evaluation criteria is present; answering a yes/no question. Qualitative scales, on the other hand, make it possible to rate the severity in an evaluation item over a given interval.

Many examples of perceptual evaluation scales can be presented such as the Frenchay Dysarthria Assessment (FDA) [8] containing both functional and perceptual evaluations or the Unified Parkinsons Disease Rating Scale (UPDRS) in which the 18th item evaluates speech on a 5-point scale.
These batteries evaluate vocal quality, phonetic realizations, prosody, respiration and intelligibility. The BECD (*Batterie d'Evaluation Clinique de la Dysarthrie* in French) [9, 10] is the most commonly-used test by clinicians for French speech. This test differentiates 35 items in order to characterize dysarthria, each item is rated on a 5-point scale.

Consequently, the use of perception for the evaluation of dysarthric speech is frequent and well documented. In addition, clinicians who evaluate the speech of patients are very well trained in detecting the phonetic characteristics associated with the physiopathology of dysarthria. However, a frequent criticism to perceptual evaluation is the subjectivity of the listeners (both naive and expert).
Several studies [11, 12] report great differences in the perceptual strategies involved in voice-quality evaluation. They point out that, in order to give their evaluation, listeners "*compare the stimulus presented to an internal standard or scale*". Obviously, many parameters may influence the distance between the stimulus and the listener's internal standard (regional accent, context, skills). Moreover, the variability between listeners' responses may be due to the signal properties they process primarily (prosody, articulation, etc.). Indeed, clinicians knowledge of habits with speech disorders and dysarthric speech production, their knowledge of the condition of the dysarthric speakers and their degree of exposure to the speakers speech alterations may influence

the evaluation results. This subjectivity associated with perceptive evaluation reduces its relevance and makes it inadequate in longitudinal studies for example. Nonetheless, evaluation by clinicians, even if subjective, is not incoherent. Most of the time, this subjectivity reflects a granularity in the perception of the degree of deviance.
An additional difficulty for speech perceptual evaluation is due to the nature of speech itself. Indeed, the production of healthy speakers is characterized by massive phonetic variations [13]. When the syntactic and semantic contexts are limited, which is often the case in perceptual experiments, variation may be difficult to interpret and may increase the ambiguity of listeners' judgment. Consequently, listeners may have difficulty in concentrating their judgment on short linguistic units due to a lack of contextual information. These limitations in speech evaluation and in human perception are difficult to fix. Nevertheless, the variation in listeners' judgment does not systematically suggest random responses and consistent results are often provided despite listeners' variable responses. In fact, the impact of variability in jurys' responses is minimized when protocols involve a large number of subjects.

Dysarthria: towards automatic approaches

To cope with these limitations, automatic approaches have rapidly emerged, as potential solutions, by providing objective tools for intelligibility assessment and anomaly detection in pathological speech [14]. In the literature, we distinguish two main approaches: those directly based on automatic speech transcription and word transcription error rate to compute an intelligibility score [15, 16], and those using automatic speech processing techniques as a means of extracting relevant information from the speech signal to perform an automatic evaluation of speech on different granularities [17, 18, 19, 20, 21]. In previous work, the authors proposed an automatic approach for phone-based anomaly detection [22] in dysarthric speech. By detecting and locating anomalies in speech production, this approach aimed to enhance manual investigation by human experts and, at the same time, to reduce the extent of their intervention by scrutinizing the speech signal. Indeed, this automatic process should make it possible to cover a significant quantity of speech production while guiding human experts towards the examination of specific parts of the speech, considered atypical. This process is notably interesting for speech production of people with mild to moderate dysarthria, for whom speech impairment may be scattered along the speech signal. Moreover, this automatic detection and location of abnormal acoustic phenomena

can have applications in clinical practice. For example, the evaluation of dysarthria by clinicians could be partially helped by a visual display of abnormal phenomena located in the speech production of dysarthric speakers, like a map. In the same manner, maps could be relevant in comparing the speech productions of a dysarthric speaker over time, during clinical treatment or rehabilitation for instance. Finally, this automatic process could be extended to other kinds of speech disorder resulting in acoustic alterations in the speech signal, such as neck or head cancers.

Motivations

In this paper, the authors investigate the behavior of the system, and in particular its potential or shortcomings, mainly over-detection of anomalies compared to a human expert. More significantly, this work attempts to tackle the question of the relationship between the human perception of alterations in speech and their modeling by automatic speech processing systems. In this context, the objective of this work is to propose an original perceptual evaluation protocol, suitable for evaluating the performance of the automatic system. This evaluation protocol aims at comparing the decisions relating to the presence of anomalies yielded by the automatic system to those of a jury composed of a large set of expert listeners (in order to minimize the effect of individual subjectivity). Both automatic and human decisions are made with regard to a selection of speech sequences produced by a large number of dysarthric patients representing four different pathologies, and by control speakers.

The rest of this article is organized as follows: section 2, the context of this research work is presented. The experimental protocol and methodology for the perceptual evaluation is presented in section 3. In section 4, the evaluation results are presented according to different aspects, while section 5 raises the question of automatic system performance and the jury's judgment tendencies. Finally, section 6 provides a conclusion and directions for future work.

2. Context: Automatic Anomaly Detection Approach and Dysarthric Corpora

2.1. Automatic Anomaly Detection Approach

The automatic phone-based anomaly detection system relies on two steps : a text-constrained phone alignment to obtain the phone segmentation and a classification of speech segments into normal and abnormal phones (anomalies).

The automatic phone segmentation of the speech utterances into phones is carried out with the help of an automatic text-constrained phone alignment tool. This tool takes into account the parameterization of the speech signals produced by a given speaker, gender-dependent acoustic models of French phones, the sequence of words pronounced by the speaker in each utterance and a phonetized, phonologically-varied lexicon of words based on a set of 37 French phones. The sequence of words comes from a manual orthographic transcription performed by a human listener following some annotation rules.

The automatic alignment process is then based on a Viterbi decoding and graph-search algorithms, the core of which is the acoustic modeling of each phone, based on a Hidden Markov Model (HMM). In this work, each phone is modeled using a 3-state context-independent HMM topology which are built using the Maximum Likelihood Estimate paradigm on the basis of about 200 hours of French radiophonic speech recordings [23]. In order to get speaker-dependent models, a three-iteration Maximum A Posteriori (MAP) adaptation is performed to adapt all the HMM parameters.

This automatic alignment process results in a couple of start and end boundaries per phone produced in the speech recordings. The precision of this automatic phone alignment was studied according to dysarthric and phonetic classes in [24].

On the basis of this alignment, a set of features considered as relevant for the anomaly detection task are extracted over each segment y_p associated with the phone p. The list of the used features can be found in [22]. These phone features are then fed into a 2-class automatic classification system based on Support Vector Machines (SVM). The SVM classification method has been largely applied in pattern recognition problems [25, 26]. Here the method is used in a 2-class classification task: the discrimination between normal and abnormal (anomaly) phones. Figure 1 describes the automatic anomaly detection process.

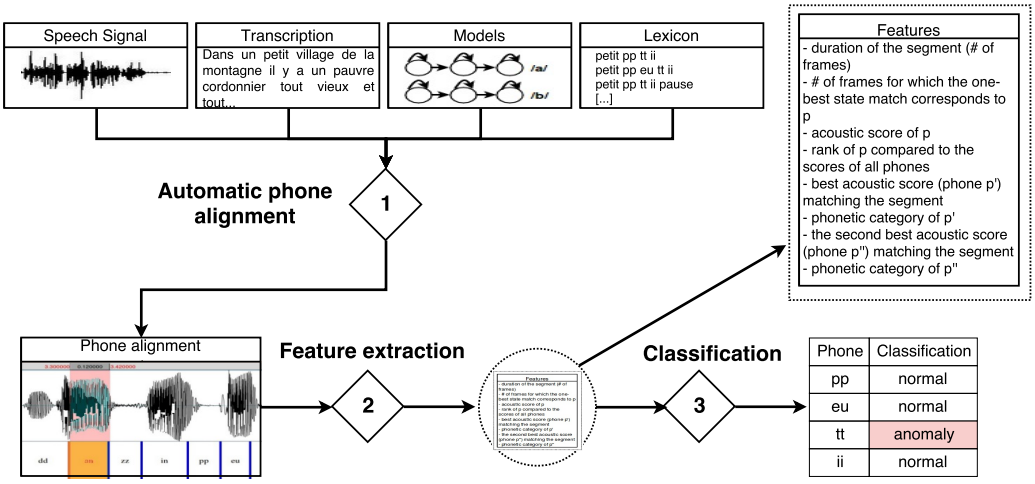

Figure 1: Automatic approach for phone-level anomaly detection.

In [22], the system was evaluated on a very limited corpus of dysarthric speech (4 female and 4 male dysarthric patients, suffering from the same pathology and 6 control speakers) annotated by one human expert. This annotation was made specially for system development and evaluation, by labeling each phone as deviant or not from an acoustic point of view. On this corpus, the system obtained a quite high averaged recall measure [1] of 0.81 (0.72 in male patients and 0.89 in female patients) and a less convincing averaged precision measure[2] of 0.63 (0.61 in male patients and 0.65 in female patients). Still in this work, the automatic system was applied on a non annotated corpus, implying a larger number of speakers (118 dysarthric patients and control speakers) and different pathologies like Parkinson's Disease (PD), Amyotrophic Lateral Sclerosis (ALS), and Cerebellar Ataxia (CA). Since no labeled data regarding anomalies was available, evaluation was carried out by observing the relationship between the rate of speech anomalies reported by the automatic system and the perceptual rates given by an expert jury on the global severity degree of dysarthria, the global degree of intelligibility and

[1]The ratio between the number of phones correctly detected as anomalies by the automatic approach and the number of zones labeled as abnormal in the reference.

[2]The ratio between the number of phones correctly detected as anomalies by the automatic approach and the total number of anomalies reported by the automatic processing (truly or falsely).

of articulation impairment, and finally, the speech rate of speakers. Analysis of the results pointed out some very interesting behavior of the automatic system, which exhibits quite relevant correlations with the majority of the perceptual criteria (e.g. between 0.8 and 0.9 for almost all of the pathologies for the global severity degree). In another work [27], the application of the automatic anomaly detection on read and spontaneous speech still highlighted the interest of such an approach.

2.2. Dysarthric speech corpora

All the selected speech sequences used in this work were extracted from French read speech recordings of the fairy tale "*Tic Tac*" (The elves and the shoemaker). In total, 40 speakers (21 male and 19 female speakers) from dysarthric speech corpora *VML* and *TypALoc* [28] were selected. Four pathologies were represented in these corpora:

- Cerebellar Ataxia (CA), caused by lesions of the cerebellum or its pathways. The cerebellum controls the balance and coordination of movements, which gives it a major role in voluntary motor control [29]. CA results in alterations in spatial-temporal organization of movement and is associated with ataxic dysarthria in the Darley classification;

- Parkinson's Disease (PD), is one of the most common degenerative neurological diseases in the world. It is linked to dysfunctions in the central nervous system, resulting from the death of cells in the substantia nigra region in the brain and, the lack of dopamine in these areas. This causes a chronic dysfunction of the central gray nuclei, essential for the execution and the control of learned motor plans [30]. The causes of this disease are still unknown and likely to include environmental, genetic, and lifestyle factors [31]. PD is associated with hypokinetic dysarthria in the Darley classification;

- Amyotrophic Lateral Sclerosis (ALS), is a primary degenerative neurological disease affecting both upper and lower motor neurons. It causes the progressive loss of motor power and has no curative treatment. Other symptoms include muscle weakness, atrophy, loss of control of all voluntary movements, postural instability and speech, phonation and swallowing difficulties. ALS is associated with a mixed dysarthria in the Darley classification;

- Lysosomal diseases (LYS) include several disorders that affect the lysosomes (entities present in each of our cells that interfere in recycling materials resulting from cellular functioning). Lysosomes fulfill their function thanks to three types of enzymes they contain and the alteration, due to genetic reasons, of this functioning results in a lysosomal disease. These diseases are often associated with mixed dysarthria. Two lysosomal disorders are represented in our corpora: Tay-Sachs and Niemann-Pick C diseases.

All the speech recordings of patients were evaluated perceptually by a jury of 11 experts who were asked to rate each patient on perceptual items of speech quality according to the GEPD evaluation protocol (a perceptive evaluation protocol containing 9 items based on the BECD evaluation protocol [9]). These items included the Dysarthria Severity Degree (DSD) rated on a scale of 0 to 3 (0 -no dysarthria, 1 -mild, 2 -moderate, 3 -severe dysarthria) and other items such as intelligibility, articulation impairment and speech rate. Three dysarthria severity degree groups were established according to the averaged perceptual evaluation issued by the jury: (1) patients with a DSD $\leq$1.5 are in severity group 1 (2) patients with DSD $\leq$2.5 are in severity group 2 (3) patients with DSD $>$ 2.5 are in severity group 3.

Table 1 details the number of patients and sequences for each pathology and their dysarthric class.

Population	Corpora	Dysarthria class	# of speakers	# of sequences
Control speakers	*TypALoc*	-	7	15
Parkinson's disease	*TypALoc*	Hypokinetic	6	15
Cerebellar ataxia	*TypALoc*	Ataxic	8	22
Amyotrophic Lateral Sclerosis	*TypALoc*	Mixed	11	28
Lysosomal storage disease	*VML*	Mixed	8	18
Total	-	-	40	98

Table 1: Information related to the corpora used for the perceptual evaluation task including the different populations and dysarthria class - control speakers and patients suffering from Parkinson's disease, cerebellar ataxia, amyotrophic lateral sclerosis, and lysosomal diseases, the number of speakers and of speech sequences per population.

3. Perceptual evaluation protocol and methodology

As mentioned above, the aim of this work was to cope with the lack of annotated corpora appropriate for evaluating the automatic detection of anomalies in speech produced by patients suffering from speech disorders, compared with normal speech. Generally, the annotation of corpora is costly and time-consuming. In our context, difficulties are increased by the fact that the automatic detection of anomalies is carried out at the phone level. A previous unpublished work we did demonstrated that the perceptual evaluation of the presence of anomalies in speech production by humans at the phone level is a very complex task, leading to very heterogeneous decisions, even when involving a large number of listeners. Based on these observations, we have proposed an original perceptual evaluation protocol of the outputs of the automatic system on the word level. The task of the jury of listeners in this protocol is still to determine the presence or not of speech deviance (anomalies), in terms of articulatory realization on sequences already annotated automatically by the approach.

3.1. Protocol design

The first feature of this protocol was to transpose the decision of the automatic system, initially at the phone level, to the word level, to facilitate the perceptual evaluation done by humans. In this way, each monosyllabic word including, at least, one phone detected as an anomaly by the automatic system was considered as abnormal. In parallel, the presence of two phones, at least, detected as anomalies in a polysyllabic word makes it abnormal. Figure 2 depicts an example of a cartography representing the automatic annotation of the production of a patient at the word level; white boxes match normal words, red boxes match abnormal words and yellow boxes represent polysyllabic words containing only one abnormal phone.

The second feature of this protocol was the set of speech sequences used for the perceptual evaluation task. Due to the cost of such tasks mentioned above, the totality of corpora automatically annotated by the system could not be used. The concentration level and cognitive effort required for each participant for the evaluation task had also to be taken into account. For these different reasons, this set of speech sequences had to be limited in size, in order to make the task feasible and efficient while relevant for the assessment of the quality of the automatic system decisions. In this way, the entire

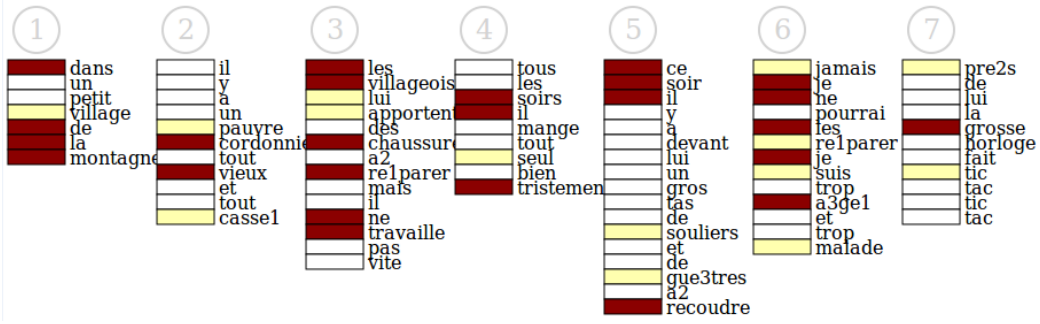

Figure 2: An example of the word-level automatically detected anomaly cartography. White boxes match normal words, red boxes match abnormal words and yellow boxes represent polysyllabic words containing only one abnormal phone.

speech corpus described in the previous section was listened to by two final-year speech therapy students [3]. Firstly, their listening made it possible to exclude recordings with low signal quality, noise or other disturbing elements. Secondly, coupled with the automatic annotation of anomalies provided by the system (figure 2), this listening highlighted different cases judged to be relevant for a finer analysis of the system behavior. Among these cases, typical errors of automatic detection systems were identified as false positives (also called false alarms), which meant that the automatic system had detected a phoneme as an anomaly whereas it was not, or false negatives, which meant that the automatic system had not detected an anomaly which was present. Contrarily, the correct detection decisions taken by the automatic system were identified as a third relevant case. Finally, in some ambiguous cases, it was possible to question and reconsider those automatically-detected anomalies, according to the listeners. Taking these considerations into account, the two speech therapists selected a limited set of sequences from the entire speech corpus and pre-classified them according to four categories for validation by the jury of experts as reported in the next section.

3.2. Speech material for perceptual evaluation

The selected sequences were extracted from the recordings using Praat [32]. Artificial silences of 400ms and 200ms were added to each at the beginning and the end respectively in order to avoid abrupt signal cuts for the

[3] These two speech therapists participated in the design of the perceptual evaluation protocol, but did not take part in the evaluation jury described later.

perceptual evaluation process.
A speech sequence contained one or several words targeted for the perceptual evaluation. For example, in the sequence "*il* ***mange*** *tout seul bien* ***tristement***" (*he eats very sadly alone*), the words "***mange***" (*eats*) and "***tristement***" (*sadly*) are targeted for the evaluation; the other words of the sequence were considered to be normally produced by the system and both speech therapists (referred to as annotators in the rest of the paper).

The different speech sequences were chosen to fit one of the following four predefined categories, regarding uniquely the target word(s) (as mentioned above, the rest of the words occurring in the speech sequences were considered as normal by the annotators and the system, independently of the categories) :

- 12.5% were referred to as "obvious segments". Here, both annotators agreed with the system annotation considering the target word(s) as abnormal; This category is rather limited in size, compared to the others since the authors were more interested by the potentially wrong behavior of the automatic system;
- 37.5% were referred to as "ambiguous segments". Here, the human annotators disagreed and were not able to decide whether the automatic annotation, considering the target word(s) as abnormal, was correct or not;
- 25% were referred to as "false negatives". Here, both annotators considered that the system failed to detect the presence of a true anomaly on the target word(s);
- 25% were referred to as "false positives". Here, both annotators considered that the system falsely labeled the target word(s) as abnormal.

Other factors shaped the set of the speech sequences. First of all, efforts had been concentrated on selecting speech produced by the largest number of patients, and representing the four pathologies available in our corpora. Secondly, efforts were made to balance the selected sequences and targeted words in order to vary their nature (grammatical, and lexical words), their length (long and short words) and their position in the sequence (start, middle, and end).

To respond to these different constraints, a total of 98 speech sequences produced by 40 speakers, including 33 dysarthric patients and 7 healthy control speakers, were finally selected for the perceptual evaluation task.

The last feature of the protocol relies on the choices of the listeners and their degree of expertise in evaluating whether or not abnormal words were present in the speech sequences. The aim of this perceptual evaluation protocol is to evaluate the quality of the outputs of an automatic system, considered itself as an "expert" since its goal is to bring some objective "expertise" to clinicians or phoneticians in their analysis of speech disorders. It seemed natural to demand that listeners, qualified in evaluating such speech disorders, participate in this evaluation protocol. A jury of expert listeners was therefore selected.

3.3. Participants

The selected jury included 29 experts aged between 22 and 58 (average age of 33). They all had French as their mother tongue and had no prior audition or learning disorders. The jury was composed of:

- 1 Ear, Nose and Throat (ENT) specialist, and speech pathologist;
- 10 speech therapists;
- 18 final-year speech therapy students.

3.4. Instructions and experimental implementation

The proposed perceptual evaluation task was then computerized using Perceval [33], an automated platform for perceptual auditory and visual tests, which can run on any Windows equipped computer. Evaluations were performed in quiet rooms and lasted between 25 and 40 minutes depending on participants. Also, each evaluation took place in one session only with no pause allowed and all participants used the same headphone set during the experiment.

The experiment was performed as follows:

1. participants were presented with an instruction list to read on the screen (see Appendix A for the translated text of the instruction list given to the listener);

2. an oral instruction was then given to all participants indicating that they should focus solely on articulatory realization and not to take prosodic or vocal aspects into account;
3. a training phase of 4 sequences was proposed in order to get the participant familiarized with the task and the use of Perceval platform;
4. the evaluation started. An orthographic transcription of the sequence appeared on the screen. Under each word, the expert had to check one of two boxes to label the word as "**normal**" or "**abnormal**". Figure 3 shows an example screen shot of the experiment. The expert could listen to each sequence up to 3 times but then had to give his/her evaluation.

It is worth noting that no information about the category the speech sequence belonged to was communicated to the experts during the perceptual evaluation. Speech sequences were presented for each listener in a totally randomized order, independently of categories.

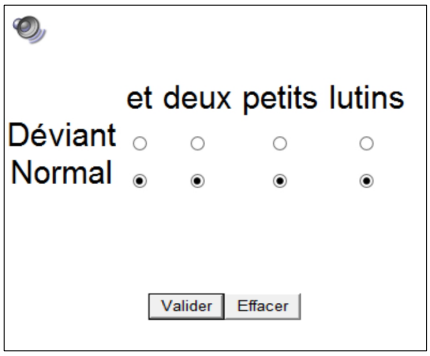

Figure 3: Screen shot from the Perceval platform used for the perceptual evaluation. The sequence tested is "et deux petits lutins" (*and two little elves*). The expert had to check one of two boxes to label each word as "**normal**" or "**déviant**" (*abnormal*) and press the "valider" (*confirm*) button.

3.5. System-Jury agreement measures

To compare the decisions made by the expert jury during their perceptual evaluation on the set of speech sequences and those of the automatic system, several agreement rates were computed:

- The *AG_targetAnomaly* rate, measuring the System-Jury agreement rate on the target words of each sequence automatically labeled as abnormal for the "obvious segments", "ambiguous segments" and "false positives" categories. This rate measures the capacity of the automatic processing in detecting present abnormal zones and how much the jury agrees with it on the detected segments. The closer to 100 the rate is, the better the automatic system detects the abnormal zones;

- the *AG_targetNormal* rate, measuring the System-Jury agreement rate on the target words of each sequence automatically labeled as normal for the "false negatives" category. This rate reflects the system inability to detect potential present anomalies (according to the two annotators). The closer to 100 the rate is, the better the automatic approach is in distinguishing anomalies from normal words and not labeling them as abnormal;

- the *AG_nonTargetNormal* rate, measuring the System-Jury agreement rate on the non-target words labeled automatically as normal for the different test sequence categories. This rate measures the system precision and capacity to distinguish between normal and abnormal words. The closer to 100 the rate is, the better the automatic approach is in not labeling normal words as anomalies;

4. Results

In this section, we present and discuss the evaluation results according to the predefined categories of the speech sequences (subsection 4.1), to the different pathologies (subsection 4.2), and to the severity degrees (subsection 4.3) present in our corpora. The results will be presented, in each case, by computing and analyzing the agreement rates presented earlier. Also, an additional analysis of the evaluation results focusing on a sub-jury of 7 selected experts is provided in subsection 4.4.

4.1. Results according to test sequence categories

4.1.1. Target abnormal words detected by the automatic approach

Figure 4 depicts the distribution of *AG_targetAnomaly* when computed for the test categories "obvious segments", "ambiguous segments" and "false positives" and *AG_targetNormal* when computed on test category "false

negatives".

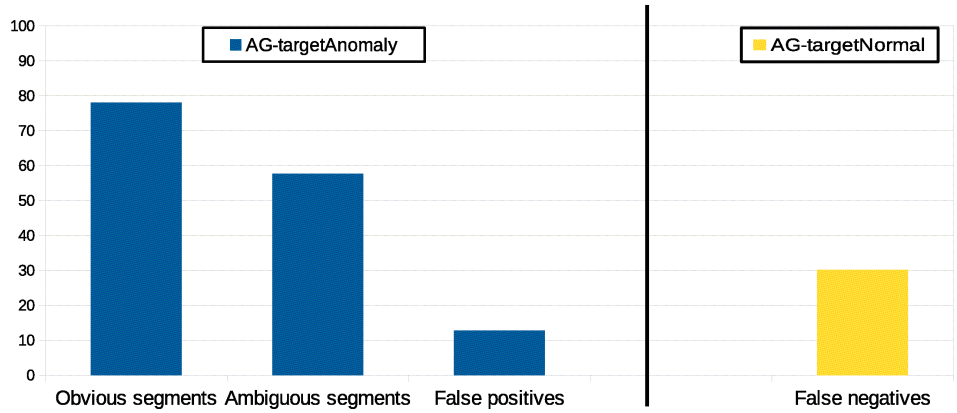

Figure 4: System-Jury *AG_targetAnomaly* and *AG_targetNormal* agreement rates (%) on automatically detected abnormal words ("obvious segments", "ambiguous segments" and "false positives") and words labeled as normal ("false negatives").

We observe a high degree of heterogeneity in the results depending on the test category reaching 78%, 58% and 13% for "obvious segments", "ambiguous segments" and "false positives" categories respectively (target words labeled as anomalies by the system).

The high *AG_targetAnomaly* rate on "obvious segments" confirms the ability of the automatic approach to detect highly distorted segments. This ability was also highlighted in [22] where the approach was able to detect 81% of phone-based anomalies annotated by an expert.

In contrast, the low *AG_targetAnomaly* rate of 13% observed on "false positives" reveals the limits of the proposed approach and its somehow approximate judgment when facing more subtle anomalies. This result calls for a more in-depth acoustical analysis of these segments in order to better comprehend the automatic system behavior and whether these segments could be related to acoustic distortions (noise, breaths, etc.) of non-pathological nature. Nevertheless, other hypotheses could also be advanced to explain this behavior such as the presence of true anomalies, which are not detected by human experts in these segments, or the presence of erroneous data in the

perceptual annotation made by an expert and used as reference to train the automatic anomaly detection system.

Considering the "ambiguous segments" test category, the System-Jury $AG\_targetAnomaly$ rate confirms the difficulty and the non trivial nature of the perceptual evaluation of dysarthric speech task even when performed by experts. Here, almost half (58%) of the jury decisions agreed with the system on the presence of an anomaly on the target words.

4.1.2. Target abnormal words undetected by the automatic approach

In this section, the focus is made on "assumed" anomalies raised by both our annotators, that the automatic system did not detect, still by observing the System-Jury agreement rate. So, considering the "false negatives" category depicted in figure 4, the System-Jury $AG\_targetNormal$ rate reaches 30%. This pretty significant rate shows that the expert jury seems neither to fully approve nor disapprove of the system behavior and the non detection of anomalies in the targeted word. Almost 1 expert in 3 agreed with the system decision.

It is worth noting that some of these test sequences contained anomalies related to word substitutions made by some speakers while reading. These substitutions were included in the manual orthographic transcription used for the text-constrained phone alignment and, therefore were taken into account by the automatic approach. However, these substitutions were not included in the perceptual evaluation process and were consequently detected by the jury of experts as anomalies. When such words are not considered in the result analysis, the $AG\_targetNormal$ reaches 36.3%.

4.1.3. Non-target normal words

Figure 5 depicts the distribution of $AG\_nonTargetNormal$ when computed on each test category.
We note that high System-Jury agreement rates on normal words are observed reaching 83%, 89%, 85% and 95% on "obvious segments", "ambiguous segments", "false negatives" and "false positives" respectively.

These rates allowed us to balance the low $AG\_targetAnomaly$ rate observed earlier in the "false positives" category. Indeed, the System-Jury overall agreement rate, $AG\_nonTargetNormal$, on automatically annotated non target normal words, across all test categories, reaches 88%. This result confirms that the automatic approach behavior is far from being arbitrary

Figure 5: System-Jury $AG\_nonTargetNormal$ agreement rates (%) on non target normal words per test sequence category.

and the observation made on the "false positives" category could be considered as marginal and restrained to a small amount of speech segments. It would be more appropriate to describe the system behavior as more severe compared to the human experts.

4.2. Inter-population variability

4.2.1. Target abnormal words detected by the automatic approach

Table 2 details the System-Jury $AG\_targetAnomaly$ rates per population and test category.

We note that the best $AG\_targetAnomaly$ rate occurred in patients suffering from lysosomal diseases (LYS) reaching 98.3% and 68.1% on "obvious segments" and "ambiguous segments" respectively. This tendency was to be expected considering that this population was involved in the modeling of the abnormal phones in the automatic system and is consistent with previous results in [22]. This does also highlight the importance of the training phase in such an automatic approach and suggests that the use of more data associated with different pathologies and dysarthric classes would improve the system performance, which is already very promising given the results reported earlier.

	Obvious segments	Ambiguous segments	False positives
CTRL	81.0	15.2	1.4
CA	71.3	59.8	9.8
PD	78.2	42.7	8.6
ALS	74.6	79.0	19.6
LYS	98.3	68.1	15.5

Table 2: System-Jury *AG_targetAnomaly* agreement rates (%) on automatically detected target abnormal words per population and test sequence category

Considering the other populations, we found that the jury members, despite their expertise level in pathological speech evaluation, were influenced by the acoustic characteristic and the overall speech quality of speakers. This is highly important considering that the instructions given to the jury explicitly restricted the evaluation task to the articulatory production of speakers. This jury's behavior is particularly observed on patients suffering from ALS for whom the jury members annotated the most anomalies compared to other populations and the *AG_targetAnomaly* rate reaches 19.6% on the "false positives" category. Indeed, the mixed dysarthria associated with this pathology is characterized by a general hypernasality, hoarseness and low speech rate. This resulted in the tendency of the jury to annotate more anomalies than expected on this population. In contrast, an opposite behavior was observed on control speakers and patients suffering from Parkinson's disease for whom the overall good quality of the speech discouraged the jury members from annotating segments as anomalies and the computed *AG_targetAnomaly* rate over the "ambiguous segments" reaches 15.2% (CTRL) and 42.7% (PD) respectively.

4.2.2. Target abnormal words undetected by the automatic approach

Table 3 details the System-Jury *AG_targetNormal* rates per population for the "false negatives" test category.
Here, it is interesting to notice that for control speakers, the jury agreed half of the time with the automatic approach whereas, for patients suffering from severe dysarthria such as ALS, CA and LYS, they tended to annotate more anomalies than the system did. This behavior suggests, once more, that the jury was highly influenced by the contextual information and the specific traits of each pathology.

	False negatives
CTRL	50.6
CA	24.9
PD	64.4
ALS	25.6
LYS	8.6

Table 3: System-Jury *AG_targetNormal* agreement rates (%) on automatically undetected abnormal words ("false negatives") per population

Also, we note that patients suffering from PD present quite singular behavior since the jury agreed with the system on the absence of an anomaly on the target words with a rate of 64.4%. This rate is even higher than the one observed in the control speakers (50.6%). This observation is somewhat singular and will require a more in-depth analysis of the set of speakers selected in both populations to explain this behavior.

4.2.3. Non-target normal words

Table 4 details the System-Jury *AG_nonTargetNormal* rates per population and test categories ("obvious segments", "ambiguous segments", "false negatives" and "false positives" test categories).

	Obvious segments	Ambiguous segments	False negatives	False positives
CTRL	99.1	99.7	97.5	100.0
CA	92.3	86.4	86.6	94.6
PD	89.1	93.7	91.2	97.3
ALS	52.9	77.1	75.3	98.0
LYS	81.9	86.6	72.9	85.6

Table 4: System-Jury *AG_nonTargetNormal* agreement rates (%) on non target normal words per population and test sequence category

We observe that for almost all populations and test sequences, the System-Jury *AG_nonTargetNormal* rates is higher than 70% reaching 100% on "false positives" sequence produced by the control speakers. In contrast,

we note that this rate is only 52.9% in non target normal words from the "obvious segments" sequences produced by ALS patients. This low result still emphasizes the importance of the contextual information for the expert jury during perceptual evaluation. Indeed, the $AG\_nonTargetNormal$ rate in ALS patients reaches 77.1% in non target normal words when surrounded by ambiguous and not prominent anomalies ("ambiguous segments") comparing to the 52.9% rate observed in non target normal words produced in an obviously deviant context ("obvious segments").

4.3. Severity variability

4.3.1. Target abnormal words detected by the automatic approach

Table 5 details the System-Jury $AG\_targetAnomaly$ rates per dysarthria severity degree group and test sequence category.

Considering the "obvious segments" and "ambiguous segments" categories, we observe that the highest computed agreement rates are observed in patients in the severity group 2 (moderate dysarthria) reaching 95.7% and 83.9% over both test categories respectively. This behavior was expected for the "obvious anomalies" test sequence category since the severity group 2 contains the most dysarthric patients used in this category (no segment produced by patients suffering from severe dysarthria were used). However, it is interesting that the agreement rate for the "ambiguous segments" for severity 3 group reaches only 51.7%.
One hypothesis to explain this behavior could be related to the dysarthria effect on all speech components. Indeed, the segments produced by patients from severity group 3 used in the "ambiguous segments" category are produced by one single patient, suffering from LYS and having an extremely low speech rate as well as articulation impairments. The automatic system could be disturbed by such speech productions, considering notably that the acoustic models used in the alignment process are based on radiophonic recordings (section 2.1), resulting in many more anomalies detected automatically. In contrast, the jury was asked to consider only articulation related anomalies and to ignore prosodic (or other) aspects. This behavior suggests that even though the design of the automatic approach did not target prosody related anomalies, such an impairment could influence it.

Considering the "false positives" test category, we observe that the $AG\_targetAnomaly$ rate reaches 31.7% for moderate dysarthria (severity group 2), but extremely low rates for the other groups of severity degree (8.0% and 1.4% for severity

group 1 and the control speaker respectively). This tendency shows that the expert jury could be highly influenced by the severity of the dysarthria and tended to 'systematically' annotate anomalies for moderate and severe dysarthria even in normal sequences and words.

	Obvious segments	Ambiguous segments	False positives
CTRL	81.0	15.2	1.4
Severity 1	71.6	47.0	8.0
Severity 2	95.7	83.9	31.7
Severity 3	-	51.7	-

Table 5: System-Jury *AG_targetAnomaly* agreement rates (%) on automatically detected target abnormal words per dysarthria severity degree and test sequence category.

4.3.2. Non-target normal words

Table 6 details the System-Jury $AG\_nonTargetAnomaly$ rates per dysarthria severity degree group and test category.
Considering all test categories, we note that higher System-Jury agreement rates are computed over control speaker and mild dysarthric patients. In fact, the higher the dysarthria is, the lower the agreement rate on normal words is (jury members annotate more anomalies). Once again, this behavior proves the subjective character of perceptual evaluation and the fact that the jury members were impacted by both the pathology and the severity of the dysarthria when evaluating each segment.

	Obvious segments	Ambiguous segments	False negatives	False positives
CTRL	99.1	99.7	97.5	100.0
Severity 1	91.4	92.2	86.7	95.9
Severity 2	49.1	78.2	71.8	80.2
Severity 3	-	48.3	-	-

Table 6: System-Jury $AG\_nonTargetNormal$ agreement rates (%) on non target normal words per dysarthria severity degree and test sequence category.

4.4. Additional analysis on a sub-jury

A more detailed analysis of listeners responses showed that the overall perception of anomalies increases from 8% to 33% depending on the listener (Figure 6). This suggests that some listeners detect few anomalies while others consider that nearly one third of the words presented in the experiment were produced with anomalies. Consequently, we raised the question whether listeners' responses were consistent or whether their subjectivity may have an influence on the results presented in the previous sections.

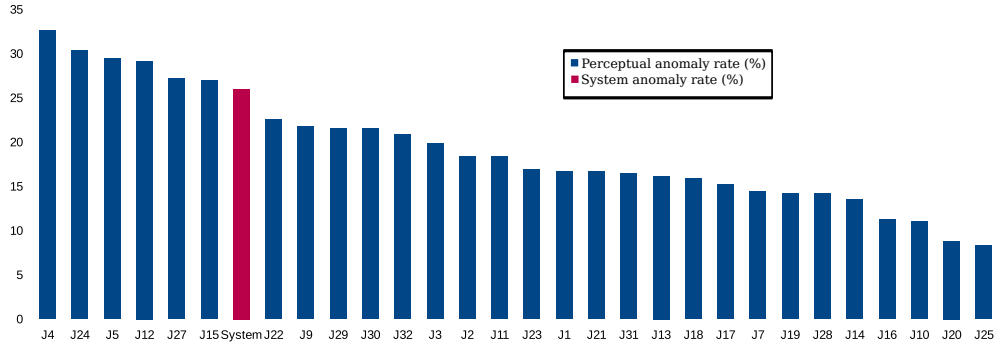

Figure 6: Anomaly rate (%) per jury member (perceptual, blue bars) and for the system (red bar).

In order to check if this jury of 29 subjects was consistent, we decided to extract a group of 7 participants from the rest of the jury (jury members J3, J5, J9, J22, J28, J30 and J32). This group, containing 2 speech therapists and 5 final-year speech therapy students, presented both higher agreement rates with the system than the rest of the jury and contained more homogeneous members in terms of annotation tendencies.

Figure 7 depicts the distribution of $AG\_targetAnomaly$, $AG\_targetNormal$ and $AG\_nonTargetNormal$ measures when computed on the different test categories ("obvious segments", "ambiguous segments", "false negatives" and "false positives") for the sub-jury.
We observed that the same pattern observed earlier on the complete jury (figure 4) is maintained for the sub-jury. Nonetheless, and considering $AG\_targetAnomaly$ measure, we note a higher System-Jury agreement rates when considering the sub-jury reaching 95.4% and 64.0% for "obvious segments" and "ambiguous segments" respectively compared to 78% and 58% for the complete jury on both test categories respectively.

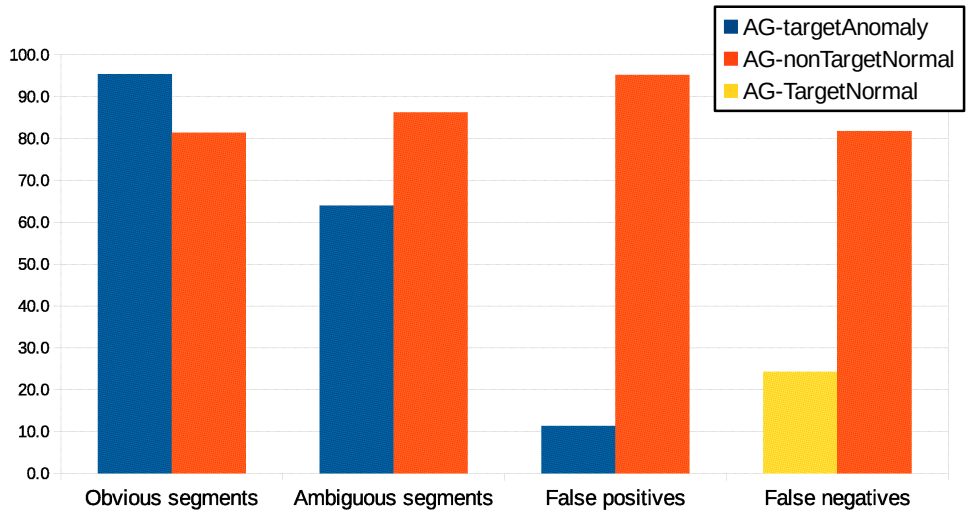

Figure 7: System-Jury agreement rates (%) on automatically detected abnormal words ($AG\_targetAnomaly$), normal target words ($AG\_targetNormal$ and normal non-target words ($AG\_nonTargetNormal$) per test sequence category for the sub-jury.

However, quite the opposite behavior was observed in the "false negatives" test category where $AG\_targetNormal$ for the sub-jury is only 24.2% whereas it reaches 30% for the complete jury. This behavior can be explained by the fact that the chosen sub-jury tended to detect more anomalies than the overall tendency which favors the system when studying anomalies detected automatically ("obvious segments", "ambiguous segments" and "false positives") but disadvantages it when considering abnormal segments that the automatic approach failed to detect ("false negatives").

Finally, comparable $AG\_nonTargetNormal$ rates were computed for both juries in the different test categories and for the sub-jury reached 81.4%, 86.2%, 81.8% and 95.1% for the "obvious segments", "ambiguous segments", "false negatives" and "false positives" categories respectively.

5. Discussion

5.1. The automatic approach quality

The results presented in section 4 confirm the capacity of the studied automatic approach in the detection of anomalies in dysarthric speech production. Also, the use of different test sequence categories made the analysis

of the approach behavior in various contexts possible. Indeed, when the anomalies are easily identified perceptually ("obvious segments"), the automatic approach has also proven its capacity in the detection of such segments (78% $AG\_targetAnomaly$).
Despite the observed severity underlined by the low $AG\_targetAnomaly$ obtained in the "false positives" category, the automatic approach is also able to demonstrate moderate and non-arbitrary behavior as supported by its high $AG\_nonTargetNormal$ reaching 88% on all test sequences.

In addition, the behavior observed concerning the "ambiguous segments" is of major interest and encouraging. Indeed, remember that in this case the presence of anomalies is much harder to identify and is more often questioned, which leads to a high variability observed among the jury members. Here, the expert jury agreed with the automatic decision in 58% of the cases "only", which is near the random threshold. Therefore, the whole jury confirmed the ambiguity of these segments. However, it is important to note that the system decisions are binary and similar to each jury's response, taken individually. The responses of the expert jury are considered to be "random" because of the comparison of the 29 responses. They are also random because the stimuli (speech sequences) are ambiguous. Indeed, the degree of phonetic deviation produced by patients is gradual and not binary. Consequently, the system should be considered an isolated jury, as reported in Figure 6 in which, that system proves to be one of the severest juries, but clearly not the severest.

From a clinical application point of view, the behavior of the automatic approach could be preferred compared with humans since it will be forced to make a binary decision (normal or abnormal) on such ambiguous speech sequences, while allowing clinicians to benefit from the reproducibility of such decisions. Furthermore, the severity of the automatic approach, potentially considered to be a limitation, yields the benefit of detecting the majority of the potential anomalies, requiring a more perceptual and acoustical in-depth analysis, which could finally be viewed as a key strength.

5.2. Inter-jury variability and judgment tendency

As expected, the jury's responses show some variation in the perception of anomalies. Nevertheless, despite this variability, the responses of listeners remain consistent and have similar tendencies. Indeed, the analysis of the responses of the sub-jury shows that a selection of homogeneous listeners

provides similar results to that of the set of 29 more heterogeneous listeners. This suggests that the ability of listeners to detect more or less anomalies does not affect the coherence of the results and the global tendencies observed in the perceptual evaluation.

Nevertheless, we probably should provide some interpretation of the variation observed in listeners' responses. This variability may be interpreted as a consequence of the difficulty of the task proposed to the jury. Indeed, listeners were asked to focus their attention on a single word which may be produced with or without an anomaly. This is not the way clinicians usually evaluate their patients. Nor is it the way humans perceive speech. The process of speech perception is more holistic and requires a large context of speech in order to evaluate if it is distorted or not. Focusing on a specific item is a very difficult task for listeners.

Moreover, listener judgments seem to be influenced by the severity of the patients. In table 2, we note that for "ambiguous segments" (for which the agreement between the experts and the system is low), the responses of the expert jury depend on the specificity of populations. For CTRL and PD speakers (low severity) the $AG\_nonTargetNormal$ is very high, while for ALS speakers (higher severity) it is lower.

Moreover, in table 6, $AG\_nonTargetNormal$ rates decreased for the severity groups 2 and 3. This means that normal speech segments (according to the automatic system and the two human annotators) produced by highly dysarthric speakers, are more frequently perceived as abnormal than segments produced by control speakers or mildly dysarthric patients due to their contextual (pathological) information. Thus, we assume that anomalies are more often detected in words when speech sequences sound pathological.

Consequently, both the subject's judgment and system anomaly detection seem to differ in two dimensions: first, the system is able to focus on short units to detect anomalies (phones, syllables, words) while subjects are accustomed to perceiving speech sounds amongst a larger linguistic context in order to take their decision; second, humans cannot avoid processing contextual information about the speakers (i.e. pathological specificity) and their decision is affected by this information, while the system does not take into account the information on speakers' specificity.

Finally, despite these differences, the results on "obvious segments" category show a strong agreement between the jury and the system. This agreement confirms the value of the automatic system and its ability to highlight speech anomalies in a clinical context. We noted that the system behavior is similar

to the most severe juries which is a positive result since the system does not miss anomalies. Consequently it provides a very useful tool to complement and refine the judgment of clinicians. Moreover, it provides useful data for acoustic studies on phoneme distortions in a pathological context.

6. Conclusion

In this paper, an original perceptual evaluation protocol of speech sequences, in the specific context of disordered speech, is proposed. Initially, the aim of this protocol and its specific design was to analyze and comprehend the behavior of an automatic anomaly detection system on dysarthric speech, by comparing the automatic annotations with those of an expert jury stemming from a perceptual evaluation. To reach this goal, speech sequences used for the perceptual evaluation were pre-classified into four different categories, mostly reflecting the unsuitable behavior of the automatic system. Different agreement rates between the expert jury's decision and the automatic system were examined according to different observation contexts.

As detailed in the paper, various results confirm the capacity as well as the relevance of the automatic approach in detecting the presence of anomalies in dysarthric speech (high $AG\_targetAnomaly$ rates on "obvious segments"). By contrast, the low $AG\_targetAnomaly$ rate computed over the "false positives" category confirms the approach tends to be more severe than human experts and requires more analysis on these segments in order to identify causes of this over-detection by the system. Experimental results also highlight that, even on the more nuanced anomalies ("ambiguous segments"), the expert jury agreed with the automatic approach decisions nearly half of the time. In this way, hypotheses motivated by the limitations recognized in the literature of the perceptual evaluation [34, 35] and by the high inter-jury variability observed during this particular evaluation were advanced to explain this behavior. Besides, this analysis reveals the main differences between the automatic and human process for the targeted detection task in terms of the size of the acoustic units and contextual information.

Unlike other fields of application, supervised automatic speech processing involved in the task of anomaly detection relies on annotated training data, for which the correctness can be questioned, as seen with the ambiguous cases this study focused on. In fact, acoustic models used for the specific anomaly detection task might be trained on doubtful annotated data. Similar remarks

could be made about the decisions of the system. Faced with some ambiguous cases, it is difficult to know whether the system responds correctly given that an expert jury may agree with it half of the time. Based on these remarks, it seems increasingly essential :

- to question the veracity of decisions taken by the expert jury, which can be used either for the model training, or for the automatic system evaluation as well as decisions taken by the automatic approach to be able to measure its performance evaluation ;
- to know how to interpret these decisions taken by both humans and the system, and the way they might interact to decide whether the system is robust enough to be used in a clinical practice for instance.

The authors also suggest raising a more primitive question : considering all the limitations of perceptual evaluation reported in the paper, should an automatic approach replicate its results and what place should be envisaged in future investigations between supervised (relying on human annotations) and semi- or unsupervised approaches for the specific task of anomaly detection, still considered a crucial step for clinicians in their evaluation of disordered speech ?

7. Acknowledgments

This work was carried out thanks to the support of the BLRI Labex (ANR-11-LABEX-0036) and the A*MIDEX project (ANR-11-IDEX-0001-02) funded by the French government Investissements dAvenir program managed by the ANR, and thanks to the French ANR projet Typaloc (ANR-12-BSH2-0003-03). We would also like to thank Laura Restivo and Laura Pianelli, who were in charge of the initial speech sequence annotation (the two annotators mentioned in the paper), of recruiting the expert jury and of organizing the perceptual evaluation-based experiments.

[1] F. L. Darley, A. E. Aronson, J. R. Brown, Clusters of deviant speech dimensions in the dysarthrias, Journal of Speech and Hearing Research 12 (1969) 462–496.

[2] B. E. Murdoch, Dysarthria: a physiological approach to assessment and treatment, 1998.

[3] J. R. Duffy, Motor speech disorders: substrates, differential diagnosis and management, Motsby- Yearbook, St Louis, 2nd edition, 2005.

[4] F. L. Darley, A. E. Aronson, J. R. Brown, Differential diagnostic patterns of dysarthria, Journal of Speech and Hearing Research 12 (1969) 246–269.

[5] F. L. Darley, A. E. Aronson, J. R. Brown, Motor speech disorders, W. B. Saunders and Co., Philadelphia, 1975.

[6] A. Lowit, R. D. Kent, Assessment of motor speech disorders, volume 1, Plural publishing, 2010.

[7] K. M. Yorkston, E. Strand, M. Kennedy, Comprehensibility of dysarthric speech: implications for assessment and treatment planning, American Journal of Speech Language Pathology 55 (1996) 55–66.

[8] P. Enderby, Frenchay dysarthric assessment, Pro-Ed, Texas (1983).

[9] P. Auzou, V. Rolland-Monnoury, Batterie d'évaluation clinique de la dysarthrie, Édition Ortho, 2006.

[10] P. Auzou, C. Ozsancak, J. R. Morris, M. Jan, F. Eustache, D. Hannequin, Voice Onset Time in aphasia, apraxia of speech and dysarthria: a review, Clinical Linguistics and Phonetics 14 (2) (2000).

[11] B. R. Gerratt, J. Kreiman, N. Antonnanzas-Barroso, G. S. Berke, Comparing internal and external standards in voice quality judgments, Journal of Speech and Hearing Research 36 (1993) 14–20.

[12] J. Kreiman, B. R. Gerratt, The perceptual structure of pathologic voice quality, The Journal of the Acoustical Society of America 100 (1996) 1787–1795.

[13] K. Johnson, Massive reduction in conversational american english, in: Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium, Tokyo, Japan: The National International Institute for Japanese Language, pp. 29–54.

[14] L. J. Ferrier, N. Jarrell, T. Carpenter, H. C. Shane, A case study of a dysarthric speaker using the Dragon Dictate voice recognition system, Journal for Computer Users in Speech and Hearing 8(1) (1992) 33–52.

[15] H. V. Sharma, M. Hasegawa-Johnson, J. Gunderson, A. Perlman, Universal access: preliminary experiments in dysarthric speech recognition, in: Proceedings of Interspeech'09, Brighton, United Kingdom.

[16] H. Christensen, S. Cunningham, C. Fox, P. Green, T. Hain, A comparative study of adaptive, automatic recognition of disordered speech, in: Proceedings of Interspeech'12, Portland, USA.

[17] C. Middag, J.-P. Martens, G. Van Nuffelen, M. De Bodt, Automated intelligibility assessment of pathological speech using phonological features, EURASIP Journal on Advances in Signal Processing 2009 (2009) 1–9.

[18] G. V. Nuffelen, C. Middag, M. D. Bodt, J.-P. Martens, Speech technology-based assessment of phoneme intelligibility in dysarthria, International journal of language and communication disorders 44(5) (2009) 716–730.

[19] T. Khan, J. Westin, M. Dougherty, Classification of speech intelligibility in parkinson's disease, Biocybernetics and Biomedical Engineering 34(1) (2014) 35–45.

[20] C. Fredouille, G. Pouchoulin, Automatic detection of abnormal zones in pathological speech, in: Intl Congress of Phonetic Sciences (ICPHs'11), Hong Kong.

[21] I. Laaridh, W. B. Kheder, C. Fredouille, C. Meunier, Automatic prediction of speech evaluation metrics for dysarthric speech, in: Proc. Interspeech, pp. 1834–1838.

[22] I. Laaridh, C. Fredouille, C. Meunier, Automatic detection of phone-based anomalies in dysarthric speech, ACM Transactions on accessible computing 6 (2015) 9:1–9:24.

[23] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, G. Gravier, ESTER phase II evaluation campaign for the rich transcription of French broadcast news, in: Proceedings of Interspeech'05, pp. 1149–1152.

[24] I. Laaridh, C. Fredouille, C. Meunier, Automatic speech processing for dysarthria: A study of inter-pathology variability, in: Proceedings of Intl Congress of Phonetic Sciences (ICPHs'15), Glasgow.

[25] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[26] B. Scholkopf, A. J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge, MA, USA, 2001.

[27] I. Laaridh, C. Fredouille, C. Meunier, Automatic anomaly detection for dysarthria across two speech styles: Read vs spontaneous speech, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portoro, Slovenia.

[28] C. Meunier, C. Fougeron, C. Fredouille, B. Bigi, L. Crevier-Buchman, E. Delais-Roussarie, L. Georgeton, A. Ghio, I. Laaridh, T. Legou, C. Pillot-Loiseau, G. Pouchoulin, The TYPALOC corpus: A collection of various dysarthric speech recordings in read and spontaneous styles, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portoro, Slovenia.

[29] C. zsancak, D. Devos, Les ataxies cérébelleuses, Les dysarthries, édition Solal Neurophysiologie et production de la parole, Part III(35) (2007) 337–348.

[30] F. Viallet, B. Teston, La dysarthrie dans la maladie de Parkinson, Les dysarthries, édition Solal Neurophysiologie et production de la parole, Part III(37) (2007) 375–382.

[31] L. Defebvre, La maladie de parkinson et les syndromes parkinsoniens, Les dysarthries, édition Solal Neurophysiologie et production de la parole, Part III(36) (2007) 364–374.

[32] P. Boersma, D. Weenink, Praat: doing phonetics by computer, http://www.praat.org/, .

[33] A. Ghio, C. André, B. Teston, C. Cavé, Perceval: une station automatisée de tests de perception et d'évaluation auditive et visuelle, Travaux

interdisciplinaires du Laboratoire parole et langage d'Aix-en-Provence (TIPA) 22 (2003) 115–133.

[34] B. J. Zyski, B. E. Weisiger, Identification of dysarthria types based on perceptual analysis, Journal of Communication Disorders 20 (1987) 367–378.

[35] S. Fex, Perceptual evaluation, Journal of voice 6 (1992) 155–158.

8. Appendix A: Instruction list

You will hear recordings of read texts in which sequences of one or more words have been extracted.
The speech produced during these readings may eventually present pathological deviations.
We ask you to judge whether the words of this sequence are deviant or not, knowing that each of the sequences can be altered in part, totally, or not at all.

You can listen to each sequence up to three times by clicking on the small speaker box.
However, if a single listening is sufficient to give your answer, you can go directly to the next sequence by clicking on the "next" box.

By default, each word appears as "normal", if one (or more) of them seems deviant, tick in the "deviant" line, the box(es) under the word(s).

Warning, the sequences have been cut from a continuous speech stream: The beginnings and/or ends can sometimes be abrupt, please do not take them into account.

The experiment should last between 30 and 40 min.
There will be a short training session to familiarize you with the task. There are no right or wrong answers, what interests us is your judgment.