



HAL
open science

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

Imed Laaridh, Christine Meunier, Corinne A Fredouille

► **To cite this version:**

Imed Laaridh, Christine Meunier, Corinne A Fredouille. Perceptual evaluation for automatic anomaly detection in disordered speech: Focus on ambiguous cases. *Speech Communication*, 2018, 105, pp.23-33. 10.1016/j.specom.2018.10.003 . hal-01959385

HAL Id: hal-01959385

<https://hal.science/hal-01959385v1>

Submitted on 18 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

2 *Dysarthria: clinical context and limitations*

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

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

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

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

37
38 Many examples of perceptual evaluation scales can be presented such as
39 the Frenchay Dysarthria Assessment (FDA) [8] containing both functional
40 and perceptual evaluations or the Unified Parkinsons Disease Rating Scale
41 (UPDRS) in which the 18th item evaluates speech on a 5-point scale.
42 These batteries evaluate vocal quality, phonetic realizations, prosody, respi-
43 ration and intelligibility. The BECD (*Batterie d'Evaluation Clinique de la*
44 *Dysarthrie* in French) [9, 10] is the most commonly-used test by clinicians
45 for French speech. This test differentiates 35 items in order to characterize
46 dysarthria, each item is rated on a 5-point scale.

47 Consequently, the use of perception for the evaluation of dysarthric speech
48 is frequent and well documented. In addition, clinicians who evaluate the
49 speech of patients are very well trained in detecting the phonetic characteris-
50 tics associated with the physiopathology of dysarthria. However, a frequent
51 criticism to perceptual evaluation is the subjectivity of the listeners (both
52 naive and expert).

53 Several studies [11, 12] report great differences in the perceptual strategies in-
54 volved in voice-quality evaluation. They point out that, in order to give their
55 evaluation, listeners ”*compare the stimulus presented to an internal standard*
56 *or scale*”. Obviously, many parameters may influence the distance between
57 the stimulus and the listener’s internal standard (regional accent, context,
58 skills). Moreover, the variability between listeners’ responses may be due
59 to the signal properties they process primarily (prosody, articulation, etc.).
60 Indeed, clinicians knowledge of habits with speech disorders and dysarthric
61 speech production, their knowledge of the condition of the dysarthric speakers
62 and their degree of exposure to the speakers speech alterations may influence

63 the evaluation results. This subjectivity associated with perceptive evalua-
64 tion reduces its relevance and makes it inadequate in longitudinal studies
65 for example. Nonetheless, evaluation by clinicians, even if subjective, is not
66 incoherent. Most of the time, this subjectivity reflects a granularity in the
67 perception of the degree of deviance.

68 An additional difficulty for speech perceptual evaluation is due to the nature
69 of speech itself. Indeed, the production of healthy speakers is character-
70 ized by massive phonetic variations [13]. When the syntactic and semantic
71 contexts are limited, which is often the case in perceptual experiments, varia-
72 tion may be difficult to interpret and may increase the ambiguity of listeners'
73 judgment. Consequently, listeners may have difficulty in concentrating their
74 judgment on short linguistic units due to a lack of contextual information.
75 These limitations in speech evaluation and in human perception are difficult
76 to fix. Nevertheless, the variation in listeners' judgment does not system-
77 atically suggest random responses and consistent results are often provided
78 despite listeners' variable responses. In fact, the impact of variability in jurors'
79 responses is minimized when protocols involve a large number of subjects.

80 *Dysarthria: towards automatic approaches*

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

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

108 *Motivations*

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

123

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

131 **2. Context: Automatic Anomaly Detection Approach and Dysarthric**
132 **Corpora**

133 *2.1. Automatic Anomaly Detection Approach*

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

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

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

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

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

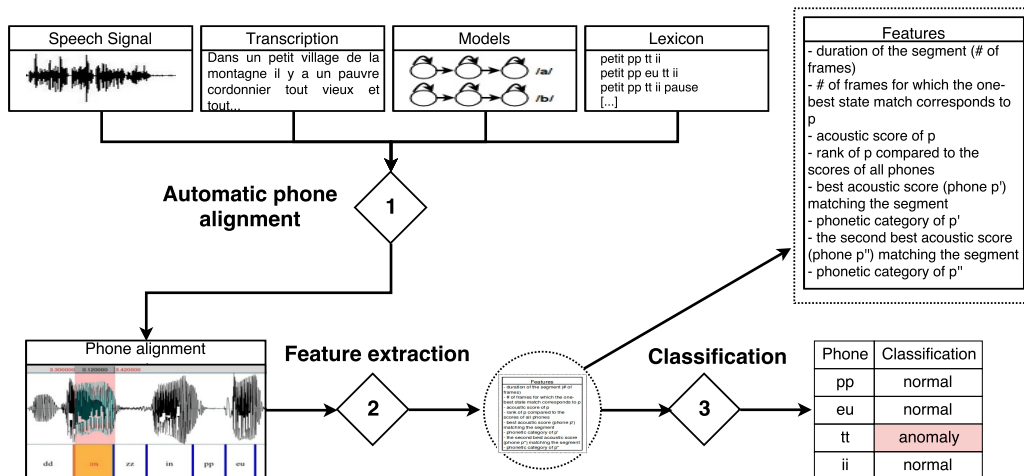


Figure 1: Automatic approach for phone-level anomaly detection.

167 In [22], the system was evaluated on a very limited corpus of dysarthric
 168 speech (4 female and 4 male dysarthric patients, suffering from the same
 169 pathology and 6 control speakers) annotated by one human expert. This
 170 annotation was made specially for system development and evaluation, by
 171 labeling each phone as deviant or not from an acoustic point of view. On
 172 this corpus, the system obtained a quite high averaged recall measure ¹ of
 173 0.81 (0.72 in male patients and 0.89 in female patients) and a less convincing
 174 averaged precision measure² of 0.63 (0.61 in male patients and 0.65 in female
 175 patients). Still in this work, the automatic system was applied on a non anno-
 176 tated corpus, implying a larger number of speakers (118 dysarthric patients
 177 and control speakers) and different pathologies like Parkinson’s Disease (PD),
 178 Amyotrophic Lateral Sclerosis (ALS), and Cerebellar Ataxia (CA). Since no
 179 labeled data regarding anomalies was available, evaluation was carried out
 180 by observing the relationship between the rate of speech anomalies reported
 181 by the automatic system and the perceptual rates given by an expert jury on
 182 the global severity degree of dysarthria, the global degree of intelligibility and

¹The 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.

²The 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).

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

190

191 2.2. Dysarthric speech corpora

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

- 197 • Cerebellar Ataxia (CA), caused by lesions of the cerebellum or its path-
198 ways. The cerebellum controls the balance and coordination of move-
199 ments, which gives it a major role in voluntary motor control [29]. CA
200 results in alterations in spatial-temporal organization of movement and
201 is associated with ataxic dysarthria in the Darley classification;
- 202 • Parkinson's Disease (PD), is one of the most common degenerative neu-
203 rological diseases in the world. It is linked to dysfunctions in the central
204 nervous system, resulting from the death of cells in the substantia ni-
205 gra region in the brain and, the lack of dopamine in these areas. This
206 causes a chronic dysfunction of the central gray nuclei, essential for
207 the execution and the control of learned motor plans [30]. The causes
208 of this disease are still unknown and likely to include environmental,
209 genetic, and lifestyle factors [31]. PD is associated with hypokinetic
210 dysarthria in the Darley classification;
- 211 • Amyotrophic Lateral Sclerosis (ALS), is a primary degenerative neuro-
212 logical disease affecting both upper and lower motor neurons. It causes
213 the progressive loss of motor power and has no curative treatment.
214 Other symptoms include muscle weakness, atrophy, loss of control of
215 all voluntary movements, postural instability and speech, phonation
216 and swallowing difficulties. ALS is associated with a mixed dysarthria
217 in the Darley classification;

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

226 All the speech recordings of patients were evaluated perceptually by a
 227 jury of 11 experts who were asked to rate each patient on perceptual items of
 228 speech quality according to the GEPD evaluation protocol (a perceptive eval-
 229 uation protocol containing 9 items based on the BECD evaluation protocol
 230 [9]). These items included the Dysarthria Severity Degree (DSD) rated on a
 231 scale of 0 to 3 (0 -no dysarthria, 1 -mild, 2 -moderate, 3 -severe dysarthria)
 232 and other items such as intelligibility, articulation impairment and speech
 233 rate. Three dysarthria severity degree groups were established according to
 234 the averaged perceptual evaluation issued by the jury: (1) patients with a
 235 DSD ≤ 1.5 are in severity group 1 (2) patients with DSD ≤ 2.5 are in severity
 236 group 2 (3) patients with DSD > 2.5 are in severity group 3.

237
 238 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	<i>TypALoc</i>	-	7	15
Parkinson’s disease	<i>TypALoc</i>	Hypokinetic	6	15
Cerebellar ataxia	<i>TypALoc</i>	Ataxic	8	22
Amyotrophic Lateral Sclerosis	<i>TypALoc</i>	Mixed	11	28
Lysosomal storage disease	<i>VML</i>	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.

240 3. Perceptual evaluation protocol and methodology

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

256 3.1. Protocol design

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

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

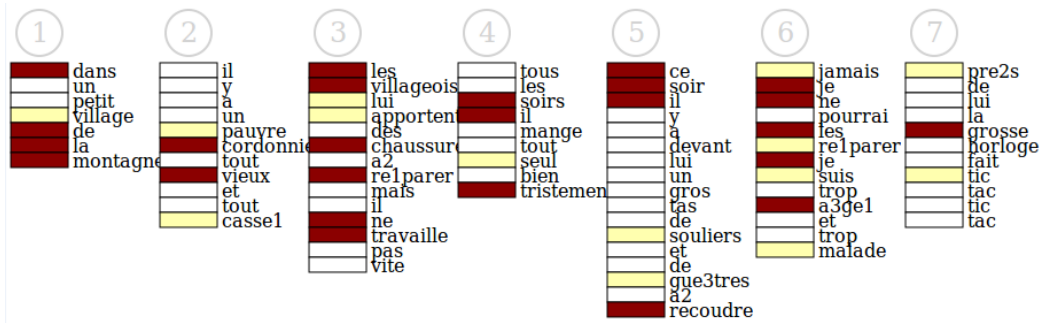


Figure 2: An example of the word-level automatically detected anomaly cartography. White boxes match normal words, red boxes match abnormal words and yellow boxes represent polysyllabic words containing only one abnormal phone.

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

292 3.2. Speech material for perceptual evaluation

293 The selected sequences were extracted from the recordings using Praat
 294 [32]. Artificial silences of 400ms and 200ms were added to each at the be-
 295 ginning and the end respectively in order to avoid abrupt signal cuts for the

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

296 perceptual evaluation process.

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

303

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

308 :

- 309 • 12.5% were referred to as "obvious segments". Here, both annotators
310 agreed with the system annotation considering the target word(s) as
311 abnormal; This category is rather limited in size, compared to the
312 others since the authors were more interested by the potentially wrong
313 behavior of the automatic system;
- 314 • 37.5% were referred to as "ambiguous segments". Here, the human
315 annotators disagreed and were not able to decide whether the automatic
316 annotation, considering the target word(s) as abnormal, was correct or
317 not;
- 318 • 25% were referred to as "false negatives". Here, both annotators con-
319 sidered that the system failed to detect the presence of a true anomaly
320 on the target word(s);
- 321 • 25% were referred to as "false positives". Here, both annotators con-
322 sidered that the system falsely labeled the target word(s) as abnormal.

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

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

333

334 The last feature of the protocol relies on the choices of the listeners and
335 their degree of expertise in evaluating whether or not abnormal words were
336 present in the speech sequences. The aim of this perceptual evaluation proto-
337 col is to evaluate the quality of the outputs of an automatic system, consid-
338 ered itself as an "expert" since its goal is to bring some objective "expertise"
339 to clinicians or phoneticians in their analysis of speech disorders. It seemed
340 natural to demand that listeners, qualified in evaluating such speech disor-
341 ders, participate in this evaluation protocol. A jury of expert listeners was
342 therefore selected.

343 *3.3. Participants*

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

- 347 • 1 Ear, Nose and Throat (ENT) specialist, and speech pathologist;
- 348 • 10 speech therapists;
- 349 • 18 final-year speech therapy students.

350 *3.4. Instructions and experimental implementation*

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

358 The experiment was performed as follows:

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

- 362 2. an oral instruction was then given to all participants indicating that
363 they should focus solely on articulatory realization and not to take
364 prosodic or vocal aspects into account;
- 365 3. a training phase of 4 sequences was proposed in order to get the par-
366 ticipant familiarized with the task and the use of Perceval platform;
- 367 4. the evaluation started. An orthographic transcription of the sequence
368 appeared on the screen. Under each word, the expert had to check one
369 of two boxes to label the word as "**normal**" or "**abnormal**". Figure
370 3 shows an example screen shot of the experiment. The expert could
371 listen to each sequence up to 3 times but then had to give his/her
372 evaluation.

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

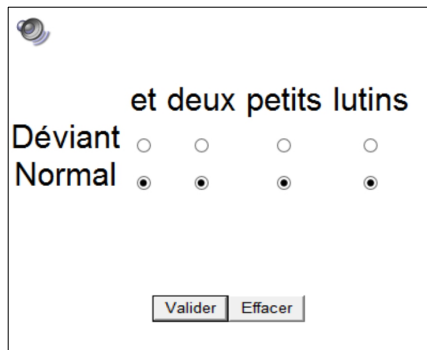


Figure 3: Screen shot from the Perceval platform used for the perceptual evaluation. The sequence tested is "et deux petits lutins" (*and two little elves*). The expert had to check one of two boxes to label each word as "**normal**" or "**déviant**" (*abnormal*) and press the "valider" (*confirm*) button.

377 3.5. System-Jury agreement measures

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

- 381 • The *AG_targetAnomaly* rate, measuring the System-Jury agreement
382 rate on the target words of each sequence automatically labeled as
383 abnormal for the "obvious segments", "ambiguous segments" and "false
384 positives" categories. This rate measures the capacity of the automatic
385 processing in detecting present abnormal zones and how much the jury
386 agrees with it on the detected segments. The closer to 100 the rate is,
387 the better the automatic system detects the abnormal zones;
- 388 • the *AG_targetNormal* rate, measuring the System-Jury agreement rate
389 on the target words of each sequence automatically labeled as normal
390 for the "false negatives" category. This rate reflects the system inability
391 to detect potential present anomalies (according to the two annotators).
392 The closer to 100 the rate is, the better the automatic approach is in
393 distinguishing anomalies from normal words and not labeling them as
394 abnormal;
- 395 • the *AG_nonTargetNormal* rate, measuring the System-Jury agree-
396 ment rate on the non-target words labeled automatically as normal
397 for the different test sequence categories. This rate measures the sys-
398 tem precision and capacity to distinguish between normal and abnormal
399 words. The closer to 100 the rate is, the better the automatic approach
400 is in not labeling normal words as anomalies;

401 4. Results

402 In this section, we present and discuss the evaluation results according
403 to the predefined categories of the speech sequences (subsection 4.1), to the
404 different pathologies (subsection 4.2), and to the severity degrees (subsection
405 4.3) present in our corpora. The results will be presented, in each case,
406 by computing and analyzing the agreement rates presented earlier. Also,
407 an additional analysis of the evaluation results focusing on a sub-jury of 7
408 selected experts is provided in subsection 4.4.

409 4.1. Results according to test sequence categories

410 4.1.1. Target abnormal words detected by the automatic approach

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

414 negatives”.

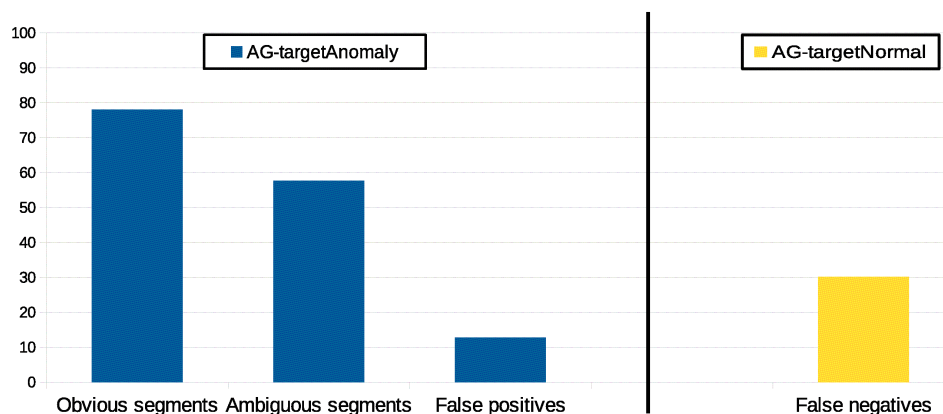


Figure 4: System-Jury *AG_targetAnomaly* and *AG_targetNormal* agreement rates (%) on automatically detected abnormal words (“obvious segments”, “ambiguous segments” and “false positives”) and words labeled as normal (“false negatives”).

415

416 We observe a high degree of heterogeneity in the results depending on
417 the test category reaching 78%, 58% and 13% for “obvious segments”, “am-
418 biguous segments” and “false positives” categories respectively (target words
419 labeled as anomalies by the system).

420 The high *AG_targetAnomaly* rate on “obvious segments” confirms the
421 ability of the automatic approach to detect highly distorted segments. This
422 ability was also highlighted in [22] where the approach was able to detect
423 81% of phone-based anomalies annotated by an expert.

424 In contrast, the low *AG_targetAnomaly* rate of 13% observed on “false pos-
425 itives” reveals the limits of the proposed approach and its somehow approx-
426 imate judgment when facing more subtle anomalies. This result calls for a
427 more in-depth acoustical analysis of these segments in order to better com-
428 prehend the automatic system behavior and whether these segments could
429 be related to acoustic distortions (noise, breaths, etc.) of non-pathological
430 nature. Nevertheless, other hypotheses could also be advanced to explain
431 this behavior such as the presence of true anomalies, which are not detected
432 by human experts in these segments, or the presence of erroneous data in the

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

435

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

441 4.1.2. Target abnormal words undetected by the automatic approach

442 In this section, the focus is made on "assumed" anomalies raised by both
443 our annotators, that the automatic system did not detect, still by observing
444 the System-Jury agreement rate. So, considering the "false negatives" cat-
445 egory depicted in figure 4, the System-Jury *AG_targetNormal* rate reaches
446 30%. This pretty significant rate shows that the expert jury seems neither to
447 fully approve nor disapprove of the system behavior and the non detection of
448 anomalies in the targeted word. Almost 1 expert in 3 agreed with the system
449 decision.

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

458 4.1.3. Non-target normal words

459 Figure 5 depicts the distribution of *AG_nonTargetNormal* when com-
460 puted on each test category.

461 We note that high System-Jury agreement rates on normal words are ob-
462 served reaching 83%, 89%, 85% and 95% on "obvious segments", "ambiguous
463 segments", "false negatives" and "false positives" respectively.

464 These rates allowed us to balance the low *AG_targetAnomaly* rate ob-
465 served earlier in the "false positives" category. Indeed, the System-Jury
466 overall agreement rate, *AG_nonTargetNormal*, on automatically annotated
467 non target normal words, across all test categories, reaches 88%. This result
468 confirms that the automatic approach behavior is far from being arbitrary

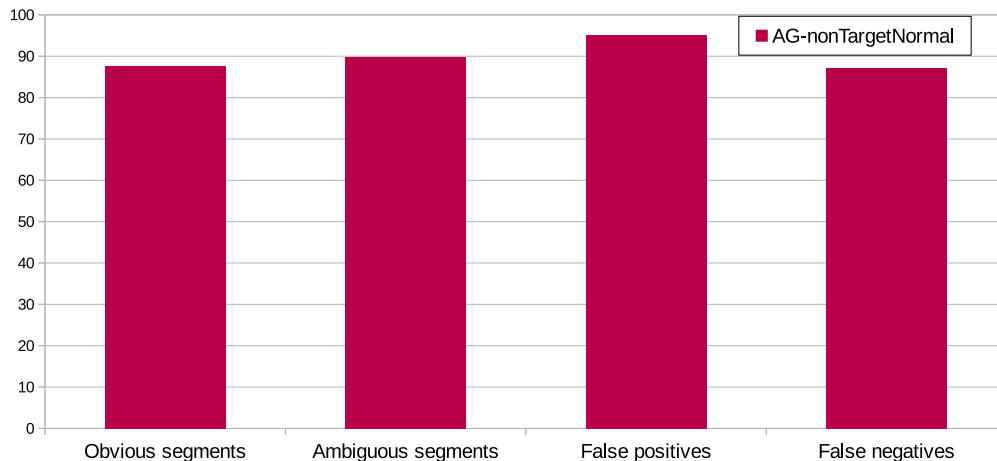


Figure 5: System-Jury *AG_nonTargetNormal* agreement rates (%) on non target normal words per test sequence category.

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

473

474 4.2. Inter-population variability

475 4.2.1. Target abnormal words detected by the automatic approach

476 Table 2 details the System-Jury *AG_targetAnomaly* rates per population
 477 and test category.

478 We note that the best *AG_targetAnomaly* rate occurred in patients suffer-
 479 ing from lysosomal diseases (LYS) reaching 98.3% and 68.1% on "obvious
 480 segments" and "ambiguous segments" respectively. This tendency was to be
 481 expected considering that this population was involved in the modeling of
 482 the abnormal phones in the automatic system and is consistent with previ-
 483 ous results in [22]. This does also highlight the importance of the training
 484 phase in such an automatic approach and suggests that the use of more data
 485 associated with different pathologies and dysarthric classes would improve
 486 the system performance, which is already very promising given the results
 487 reported earlier.

488

	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

489 Considering the other populations, we found that the jury members, de-
490 spite their expertise level in pathological speech evaluation, were influenced
491 by the acoustic characteristic and the overall speech quality of speakers. This
492 is highly important considering that the instructions given to the jury explic-
493 itly restricted the evaluation task to the articulatory production of speakers.
494 This jury’s behavior is particularly observed on patients suffering from ALS
495 for whom the jury members annotated the most anomalies compared to other
496 populations and the *AG_targetAnomaly* rate reaches 19.6% on the ”false pos-
497 itives” category. Indeed, the mixed dysarthria associated with this pathology
498 is characterized by a general hypernasality, hoarseness and low speech rate.
499 This resulted in the tendency of the jury to annotate more anomalies than
500 expected on this population. In contrast, an opposite behavior was observed
501 on control speakers and patients suffering from Parkinson’s disease for whom
502 the overall good quality of the speech discouraged the jury members from
503 annotating segments as anomalies and the computed *AG_targetAnomaly*
504 rate over the ”ambiguous segments” reaches 15.2% (CTRL) and 42.7% (PD)
505 respectively.

506 4.2.2. Target abnormal words undetected by the automatic approach

507 Table 3 details the System-Jury *AG_targetNormal* rates per population
508 for the ”false negatives” test category.
509 Here, it is interesting to notice that for control speakers, the jury agreed half
510 of the time with the automatic approach whereas, for patients suffering from
511 severe dysarthria such as ALS, CA and LYS, they tended to annotate more
512 anomalies than the system did. This behavior suggests, once more, that the
513 jury was highly influenced by the contextual information and the specific
514 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

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

522 4.2.3. Non-target normal words

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

526

	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

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

530 we note that this rate is only 52.9% in non target normal words from the
531 "obvious segments" sequences produced by ALS patients. This low result
532 still emphasizes the importance of the contextual information for the expert
533 jury during perceptual evaluation. Indeed, the *AG_nonTargetNormal* rate
534 in ALS patients reaches 77.1% in non target normal words when surrounded
535 by ambiguous and not prominent anomalies ("ambiguous segments") com-
536 paring to the 52.9% rate observed in non target normal words produced in
537 an obviously deviant context ("obvious segments").

538 4.3. Severity variability

539 4.3.1. Target abnormal words detected by the automatic approach

540 Table 5 details the System-Jury *AG_targetAnomaly* rates per dysarthria
541 severity degree group and test sequence category.

542
543 Considering the "obvious segments" and "ambiguous segments" cate-
544 gories, we observe that the highest computed agreement rates are observed
545 in patients in the severity group 2 (moderate dysarthria) reaching 95.7% and
546 83.9% over both test categories respectively. This behavior was expected
547 for the "obvious anomalies" test sequence category since the severity group
548 2 contains the most dysarthric patients used in this category (no segment
549 produced by patients suffering from severe dysarthria were used). However,
550 it is interesting that the agreement rate for the "ambiguous segments" for
551 severity 3 group reaches only 51.7%.

552 One hypothesis to explain this behavior could be related to the dysarthria
553 effect on all speech components. Indeed, the segments produced by patients
554 from severity group 3 used in the "ambiguous segments" category are pro-
555 duced by one single patient, suffering from LYS and having an extremely low
556 speech rate as well as articulation impairments. The automatic system could
557 be disturbed by such speech productions, considering notably that the acous-
558 tic models used in the alignment process are based on radiophonic recordings
559 (section 2.1), resulting in many more anomalies detected automatically. In
560 contrast, the jury was asked to consider only articulation related anomalies
561 and to ignore prosodic (or other) aspects. This behavior suggests that even
562 though the design of the automatic approach did not target prosody related
563 anomalies, such an impairment could influence it.

564 Considering the "false positives" test category, we observe that the *AG_targetAnomaly*
565 rate reaches 31.7% for moderate dysarthria (severity group 2), but extremely
566 low rates for the other groups of severity degree (8.0% and 1.4% for severity

567 group 1 and the control speaker respectively). This tendency shows that
 568 the expert jury could be highly influenced by the severity of the dysarthria
 569 and tended to 'systematically' annotate anomalies for moderate and severe
 570 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.

571 4.3.2. Non-target normal words

572 Table 6 details the System-Jury *AG_nonTargetAnomaly* rates per dysarthria
 573 severity degree group and test category.

574 Considering all test categories, we note that higher System-Jury agreement
 575 rates are computed over control speaker and mild dysarthric patients. In
 576 fact, the higher the dysarthria is, the lower the agreement rate on normal
 577 words is (jury members annotate more anomalies). Once again, this behavior
 578 proves the subjective character of perceptual evaluation and the fact that the
 579 jury members were impacted by both the pathology and the severity of the
 580 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.

581 4.4. Additional analysis on a sub-jury

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

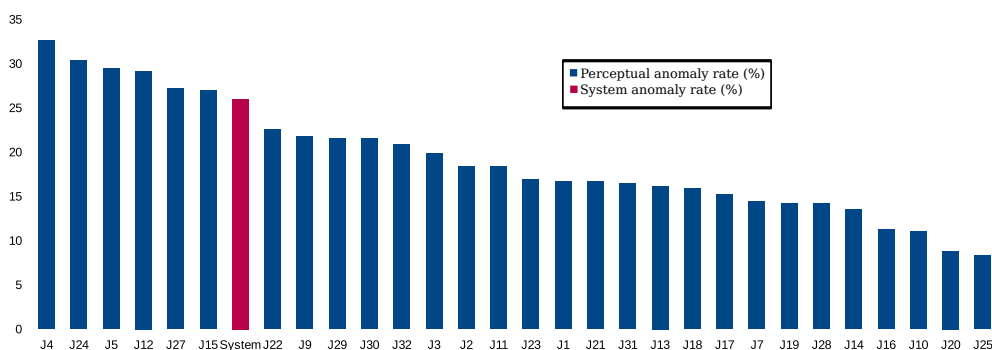


Figure 6: Anomaly rate (%) per jury member (perceptual, blue bars) and for the system (red bar).

589 In order to check if this jury of 29 subjects was consistent, we decided to
 590 extract a group of 7 participants from the rest of the jury (jury members J3,
 591 J5, J9, J22, J28, J30 and J32). This group, containing 2 speech therapists and
 592 5 final-year speech therapy students, presented both higher agreement rates
 593 with the system than the rest of the jury and contained more homogeneous
 594 members in terms of annotation tendencies.

595 Figure 7 depicts the distribution of *AG_targetAnomaly*, *AG_targetNormal*
 596 and *AG_nonTargetNormal* measures when computed on the different test
 597 categories (“obvious segments”, “ambiguous segments”, “false negatives” and
 598 “false positives”) for the sub-jury.

599 We observed that the same pattern observed earlier on the complete jury (fig-
 600 ure 4) is maintained for the sub-jury. Nonetheless, and considering *AG_targetAnomaly*
 601 measure, we note a higher System-Jury agreement rates when considering the
 602 sub-jury reaching 95.4% and 64.0% for “obvious segments” and “ambiguous
 603 segments” respectively compared to 78% and 58% for the complete jury on
 604 both test categories respectively.

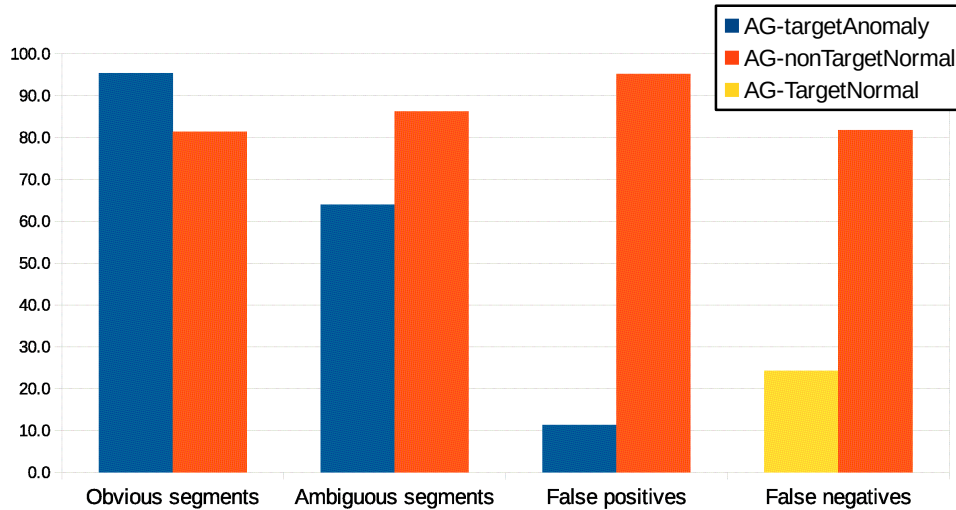


Figure 7: System-Jury agreement rates (%) on automatically detected abnormal words (*AG_targetAnomaly*), normal target words (*AG_targetNormal*) and normal non-target words (*AG_nonTargetNormal*) per test sequence category for the sub-jury.

605 However, quite the opposite behavior was observed in the "false nega-
 606 tives" test category where *AG_targetNormal* for the sub-jury is only 24.2%
 607 whereas it reaches 30% for the complete jury. This behavior can be explained
 608 by the fact that the chosen sub-jury tended to detect more anomalies than
 609 the overall tendency which favors the system when studying anomalies de-
 610 tected automatically ("obvious segments", "ambiguous segments" and "false
 611 positives") but disadvantages it when considering abnormal segments that
 612 the automatic approach failed to detect ("false negatives").

613 Finally, comparable *AG_nonTargetNormal* rates were computed for both
 614 juries in the different test categories and for the sub-jury reached 81.4%,
 615 86.2%, 81.8% and 95.1% for the "obvious segments", "ambiguous segments",
 616 "false negatives" and "false positives" categories respectively.

617 5. Discussion

618 5.1. The automatic approach quality

619 The results presented in section 4 confirm the capacity of the studied
 620 automatic approach in the detection of anomalies in dysarthric speech pro-
 621 duction. Also, the use of different test sequence categories made the analysis

622 of the approach behavior in various contexts possible. Indeed, when the
623 anomalies are easily identified perceptually ("obvious segments"), the auto-
624 matic approach has also proven its capacity in the detection of such segments
625 (78% *AG_targetAnomaly*).

626 Despite the observed severity underlined by the low *AG_targetAnomaly* ob-
627 tained in the "false positives" category, the automatic approach is also able
628 to demonstrate moderate and non-arbitrary behavior as supported by its
629 high *AG_nonTargetNormal* reaching 88% on all test sequences.

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

645
646 From a clinical application point of view, the behavior of the automatic
647 approach could be preferred compared with humans since it will be forced
648 to make a binary decision (normal or abnormal) on such ambiguous speech
649 sequences, while allowing clinicians to benefit from the reproducibility of such
650 decisions. Furthermore, the severity of the automatic approach, potentially
651 considered to be a limitation, yields the benefit of detecting the majority of
652 the potential anomalies, requiring a more perceptual and acoustical in-depth
653 analysis, which could finally be viewed as a key strength.

654 5.2. *Inter-jury variability and judgment tendency*

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

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

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

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

678 Moreover, in table 6, *AG_nonTargetNormal* rates decreased for the sever-
679 ity groups 2 and 3. This means that normal speech segments (according to
680 the automatic system and the two human annotators) produced by highly
681 dysarthric speakers, are more frequently perceived as abnormal than seg-
682 ments produced by control speakers or mildly dysarthric patients due to their
683 contextual (pathological) information. Thus, we assume that anomalies are
684 more often detected in words when speech sequences sound pathological.

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

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

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

701 6. Conclusion

702 In this paper, an original perceptual evaluation protocol of speech se-
703 quences, in the specific context of disordered speech, is proposed. Initially,
704 the aim of this protocol and its specific design was to analyze and compre-
705 hend the behavior of an automatic anomaly detection system on dysarthric
706 speech, by comparing the automatic annotations with those of an expert jury
707 stemming from a perceptual evaluation. To reach this goal, speech sequences
708 used for the perceptual evaluation were pre-classified into four different cat-
709 egories, mostly reflecting the unsuitable behavior of the automatic system.
710 Different agreement rates between the expert jury’s decision and the auto-
711 matic system were examined according to different observation contexts.

712 As detailed in the paper, various results confirm the capacity as well as the
713 relevance of the automatic approach in detecting the presence of anomalies
714 in dysarthric speech (high *AG_targetAnomaly* rates on ”obvious segments”).
715 By contrast, the low *AG_targetAnomaly* rate computed over the ”false posi-
716 tives” category confirms the approach tends to be more severe than human
717 experts and requires more analysis on these segments in order to identify
718 causes of this over-detection by the system. Experimental results also high-
719 light that, even on the more nuanced anomalies (”ambiguous segments”), the
720 expert jury agreed with the automatic approach decisions nearly half of the
721 time. In this way, hypotheses motivated by the limitations recognized in the
722 literature of the perceptual evaluation [34, 35] and by the high inter-jury vari-
723 ability observed during this particular evaluation were advanced to explain
724 this behavior. Besides, this analysis reveals the main differences between the
725 automatic and human process for the targeted detection task in terms of the
726 size of the acoustic units and contextual information.

727
728 Unlike other fields of application, supervised automatic speech processing
729 involved in the task of anomaly detection relies on annotated training data,
730 for which the correctness can be questioned, as seen with the ambiguous cases
731 this study focused on. In fact, acoustic models used for the specific anomaly
732 detection task might be trained on doubtful annotated data. Similar remarks

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

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

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

751 **7. Acknowledgments**

752 This work was carried out thanks to the support of the BLRI Labex
753 (ANR-11-LABEX-0036) and the A*MIDEX project (ANR-11-IDEX-0001-
754 02) funded by the French government Investissements dAvenir program man-
755 aged by the ANR, and thanks to the French ANR projet Typaloc (ANR-
756 12-BSH2-0003-03). We would also like to thank Laura Restivo and Laura
757 Pianelli, who were in charge of the initial speech sequence annotation (the
758 two annotators mentioned in the paper), of recruiting the expert jury and of
759 organizing the perceptual evaluation-based experiments.

- 760 [1] F. L. Darley, A. E. Aronson, J. R. Brown, Clusters of deviant speech
761 dimensions in the dysarthrias, *Journal of Speech and Hearing Research*
762 12 (1969) 462–496.
- 763 [2] B. E. Murdoch, *Dysarthria: a physiological approach to assessment and*
764 *treatment*, 1998.

- 765 [3] J. R. Duffy, Motor speech disorders: substrates, differential diagnosis
766 and management, Motsby- Yearbook, St Louis, 2nd edition, 2005.
- 767 [4] F. L. Darley, A. E. Aronson, J. R. Brown, Differential diagnostic pat-
768 terns of dysarthria, *Journal of Speech and Hearing Research* 12 (1969)
769 246–269.
- 770 [5] F. L. Darley, A. E. Aronson, J. R. Brown, Motor speech disorders, W.
771 B. Saunders and Co., Philadelphia, 1975.
- 772 [6] A. Lowit, R. D. Kent, Assessment of motor speech disorders, volume 1,
773 Plural publishing, 2010.
- 774 [7] K. M. Yorkston, E. Strand, M. Kennedy, Comprehensibility of
775 dysarthric speech: implications for assessment and treatment planning,
776 *American Journal of Speech Language Pathology* 55 (1996) 55–66.
- 777 [8] P. Enderby, Frenchay dysarthric assessment, Pro-Ed, Texas (1983).
- 778 [9] P. Auzou, V. Rolland-Monnoury, Batterie d'évaluation clinique de la
779 dysarthrie, Édition Ortho, 2006.
- 780 [10] P. Auzou, C. Ozsancak, J. R. Morris, M. Jan, F. Eustache, D. Han-
781 nequin, Voice Onset Time in aphasia, apraxia of speech and dysarthria:
782 a review, *Clinical Linguistics and Phonetics* 14 (2) (2000).
- 783 [11] B. R. Gerratt, J. Kreiman, N. Antonanzas-Barroso, G. S. Berke, Com-
784 paring internal and external standards in voice quality judgments, *Jour-
785 nal of Speech and Hearing Research* 36 (1993) 14–20.
- 786 [12] J. Kreiman, B. R. Gerratt, The perceptual structure of pathologic voice
787 quality, *The Journal of the Acoustical Society of America* 100 (1996)
788 1787–1795.
- 789 [13] K. Johnson, Massive reduction in conversational american english, in:
790 Spontaneous speech: Data and analysis. Proceedings of the 1st session
791 of the 10th international symposium, Tokyo, Japan: The National In-
792 ternational Institute for Japanese Language, pp. 29–54.
- 793 [14] L. J. Ferrier, N. Jarrell, T. Carpenter, H. C. Shane, A case study of a
794 dysarthric speaker using the Dragon Dictate voice recognition system,
795 *Journal for Computer Users in Speech and Hearing* 8(1) (1992) 33–52.

- 796 [15] H. V. Sharma, M. Hasegawa-Johnson, J. Gunderson, A. Perlman, Uni-
797 versal access: preliminary experiments in dysarthric speech recognition,
798 in: Proceedings of Interspeech'09, Brighton, United Kingdom.
- 799 [16] H. Christensen, S. Cunningham, C. Fox, P. Green, T. Hain, A compar-
800 ative study of adaptive, automatic recognition of disordered speech, in:
801 Proceedings of Interspeech'12, Portland, USA.
- 802 [17] C. Middag, J.-P. Martens, G. Van Nuffelen, M. De Bodt, Automated
803 intelligibility assessment of pathological speech using phonological fea-
804 tures, EURASIP Journal on Advances in Signal Processing 2009 (2009)
805 1–9.
- 806 [18] G. V. Nuffelen, C. Middag, M. D. Bodt, J.-P. Martens, Speech
807 technology-based assessment of phoneme intelligibility in dysarthria,
808 International journal of language and communication disorders 44(5)
809 (2009) 716–730.
- 810 [19] T. Khan, J. Westin, M. Dougherty, Classification of speech intelligibility
811 in parkinson's disease, Biocybernetics and Biomedical Engineering 34(1)
812 (2014) 35–45.
- 813 [20] C. Fredouille, G. Pouchoulin, Automatic detection of abnormal zones in
814 pathological speech, in: Intl Congress of Phonetic Sciences (ICPHs'11),
815 Hong Kong.
- 816 [21] I. Laaridh, W. B. Kheder, C. Fredouille, C. Meunier, Automatic pre-
817 diction of speech evaluation metrics for dysarthric speech, in: Proc.
818 Interspeech, pp. 1834–1838.
- 819 [22] I. Laaridh, C. Fredouille, C. Meunier, Automatic detection of phone-
820 based anomalies in dysarthric speech, ACM Transactions on accessible
821 computing 6 (2015) 9:1–9:24.
- 822 [23] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre,
823 G. Gravier, ESTER phase II evaluation campaign for the rich tran-
824 scription of French broadcast news, in: Proceedings of Interspeech'05,
825 pp. 1149–1152.

- 826 [24] I. Laaridh, C. Fredouille, C. Meunier, Automatic speech processing for
827 dysarthria: A study of inter-pathology variability, in: Proceedings of
828 Intl Congress of Phonetic Sciences (ICPHs'15), Glasgow.
- 829 [25] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag
830 New York, Inc., New York, NY, USA, 1995.
- 831 [26] B. Scholkopf, A. J. Smola, Learning with Kernels: Support Vector Ma-
832 chines, Regularization, Optimization, and Beyond, MIT Press, Cam-
833 bridge, MA, USA, 2001.
- 834 [27] I. Laaridh, C. Fredouille, C. Meunier, Automatic anomaly detection
835 for dysarthria across two speech styles: Read vs spontaneous speech,
836 in: Proceedings of the Tenth International Conference on Language Re-
837 sources and Evaluation (LREC 2016), Portoro, Slovenia.
- 838 [28] C. Meunier, C. Fougeron, C. Fredouille, B. Bigi, L. Crevier-Buchman,
839 E. Delais-Roussarie, L. Georgeton, A. Ghio, I. Laaridh, T. Legou,
840 C. Pillot-Loiseau, G. Pouchoulin, The TYPALOC corpus: A collec-
841 tion of various dysarthric speech recordings in read and spontaneous
842 styles, in: Proceedings of the Tenth International Conference on Lan-
843 guage Resources and Evaluation (LREC'16), Portoro, Slovenia.
- 844 [29] C. zsanca, D. Devos, Les ataxies cérébelleuses, Les dysarthries, édition
845 Solal Neurophysiologie et production de la parole, Part III(35) (2007)
846 337–348.
- 847 [30] F. Viallet, B. Teston, La dysarthrie dans la maladie de Parkinson, Les
848 dysarthries, édition Solal Neurophysiologie et production de la parole,
849 Part III(37) (2007) 375–382.
- 850 [31] L. Defebvre, La maladie de parkinson et les syndromes parkinsoniens,
851 Les dysarthries, édition Solal Neurophysiologie et production de la pa-
852 role, Part III(36) (2007) 364–374.
- 853 [32] P. Boersma, D. Weenink, Praat: doing phonetics by computer,
854 <http://www.praat.org/>, .
- 855 [33] A. Ghio, C. André, B. Teston, C. Cavé, Perceval: une station automa-
856 tisée de tests de perception et d'évaluation auditive et visuelle, Travaux

- 857 interdisciplinaires du Laboratoire parole et langage d'Aix-en-Provence
858 (TIPA) 22 (2003) 115–133.
- 859 [34] B. J. Zyski, B. E. Weisiger, Identification of dysarthria types based
860 on perceptual analysis, *Journal of Communication Disorders* 20 (1987)
861 367–378.
- 862 [35] S. Fex, Perceptual evaluation, *Journal of voice* 6 (1992) 155–158.

863 **8. Appendix A: Instruction list**

864 You will hear recordings of read texts in which sequences of one or more
865 words have been extracted.

866 The speech produced during these readings may eventually present patho-
867 logical deviations.

868 We ask you to judge whether the words of this sequence are deviant or not,
869 knowing that each of the sequences can be altered in part, totally, or not at
870 all.

871

872 You can listen to each sequence up to three times by clicking on the small
873 speaker box.

874 However, if a single listening is sufficient to give your answer, you can go
875 directly to the next sequence by clicking on the "next" box.

876

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

879

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

883

884 The experiment should last between 30 and 40 min.

885 There will be a short training session to familiarize you with the task. There
886 are no right or wrong answers, what interests us is your judgment.