Dysarthric speech evaluation: automatic and perceptual approaches

Imed Laaridh[1,2], Christine Meunier[2], Corinne Fredouille[1]

[1]University of Avignon, CERI/LIA, France

[2]Aix Marseille Université, CNRS, LPL UMR 7309, 13100, Aix-en-Provence, France

imed.laaridh@alumni.univ-avignon.fr, christine.meunier@lpl-aix.fr, corinne.fredouille@univ-avignon.fr

Abstract

Perceptual evaluation is still the most common method in clinical practice for the diagnosis and monitoring of 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). Clinicians have, therefore, expressed their need for new objective evaluation tools more adapted to longitudinal studies or rehabilitation context.

We proposed earlier an automatic approach for the anomaly detection at the phone level for dysarthric speech. The system behavior was studied and validated on different corpora and speech styles. Nonetheless, the lack of annotated French dysarthric speech corpora has limited our capacity to analyze some aspects of its behavior, and notably,its severity (more anomalies detected automatically compared with human expert). To overcome this limitation, we proposed an original perceptual evaluation protocol applied to a limited set of decisions made by the automatic system, related to the presence of anomalies. This evaluation was carried out by a jury of 29 non-naive individuals. In addition to interesting information related to the system behavior, the evaluation protocol highlighted the difficulty for a human, even expert, to apprehend and detect deviations at the word level in dysarthric speech.

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

1. Introduction

Dysarthria is a motor speech disorder that is a consequence of neurological damage located in the central or/and the peripheral nervous system. This may result in disturbances in any of the components involved in speech production and may be reflected by weakness, spasticity, incoordination, involuntary movements, or abnormal muscle tone depending on the location of the neurological damage (Duffy, 2005).

Perceptual evaluation by a set of listeners is the most common paradigm used to evaluate the characteristics and severity of impairment in speech pathologies. Furthermore, the current classification of dysarthria into six categories is based on a perceptual description of deviant speech dimensions. ((Darley et al., 1969b; Darley et al., 1969a; Darley et al., 1975). The clinical evaluation of patients is also based on several batteries of tests in which the production of dysarthric speakers is coded perceptually by clinicians. These batteries evaluate the vocal quality, phonetic realizations, prosody, respiration and intelligibility. The BECD (*Batterie d'Evaluation Clinique de la Dysarthrie*) (Auzou and Rolland-Monnoury, 2006) is the most commonly used test by clinicians for French speech. This test differentiates 35 items in order to characterize dysarthria. Consequently, the use of perception for the evaluation of dysarthric speech is frequent and well documented. And the clinicians evaluating the speech of patients are very well trained to 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).

In some previous work, the authors have proposed an automatic phone-based anomaly detection approach (Laaridh et al., 2015a). By detecting and localizing anomalies in speech production, this approach aims at enhancing the manual investigation of human experts and, at the same time, reducing the extent of their intervention by scrutinizing the speech signal. Indeed, this automatic process should permit treating a larger amount of speech production while guiding human experts to focus on specific parts of the speech, considered as atypical. This process is notably interesting for speech productions of people with mild to moderate dysarthria, for which speech impairment may be scattered along the speech signal. Moreover, this automatic detection and localization 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 localized in the speech production of dysarthric speakers, like a map. In the same manner, maps could be relevant to compare the speech productions of a dysarthric speaker over time, in clinical treatment or rehabilitation for instance. Finally, this automatic process could be extended to other kinds of speech disorders resulting in acoustic alterations in the speech signal, such as larynx or head and neck cancers.

In this paper, the authors investigate on the behavior of the system, and typically on its potential quality or shortcoming to over-detect anomalies compared to a human expert. The larger research question this work tries to handle could be also that 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 is based on the comparison of the output decisions yielded by the system relating to the presence of anomalies, to those of a jury composed of a large set of expert listeners (in order to minimize the effect of individual subjectivity) on a selection of speech sequences produced by a large number of dysarthric patients representing four different pathologies, and control

speakers.

The rest of this article is organized as follows. In section 2, the automatic anomaly detection approach and the motivations of this research work are presented. The experimental protocol and corpora used for the perceptual evaluation are presented in section 3. In section 4, the evaluation results are presented and discussed according to different aspects. Finally, section 5 provides a conclusion and directions for future work.

2. Motivations

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 thanks to an automatic text-constrained phone alignment tool (Laaridh et al., 2015b). This tool takes as input the parameterization of the speech signal 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. Then, 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 features used can be found in (Laaridh et al., 2015a). These phone features are then fed into a 2-class automatic classification system based on Support Vector Machines (SVM) (Vapnik, 1995; Scholkopf and Smola, 2001). Figure 1 describes the automatic anomaly detection process.

In (Laaridh et al., 2015a), 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 especially for the 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 on male patients and 0.89 on female patients) and a less convincing averaged precision measure[2] of 0.63 (0.61 on male patients and 0.65 on 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 normal speakers) and different pathologies like Parkinson's Disease, Amyotrophic Lateral Sclerosis, and Cerebellar Ataxia. Since no labeled data regarding anomalies was available, evaluation was carried out by observing the relationship between the

amount 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 of articulation impairment, and finally, the speech rate of speakers. Analysis of the results pointed out very interesting behavior of the automatic system, which exhibits very 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 (Laaridh et al., 2016), the application of the automatic anomaly detection on read and spontaneous speech still highlighted the interest of such an approach.

2.2. Proposed protocol features

The aim of the work reported in this paper is to cope with the lack of annotated corpora suitable 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 long, costly and time-consuming. In our context, difficulties increase by the fact that the automatic detection of anomalies is carried out at the phone level. Previous unpublished work we did demonstrated that the perceptual evaluation of the presence of anomalies in speech production by human at the phone level is a very complex task, leading to very heterogeneous decisions, even while involving a large number of listeners. Based on these observations, we propose an original perceptual evaluation protocol of the outputs of the automatic system. The task of listeners in this protocol is still to determine the presence of speech deviance (anomalies), in terms of articulatory realization.

The first feature of this protocol is 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 is considered as abnormal. In parallel, the presence of two phones, at least, detected as an anomaly in a polysyllabic word makes it abnormal.

The second feature of this protocol is the set of speech sequences used for the perceptual evaluation task. Due to the cost of such tasks mentioned above, the entire corpora automatically annotated by the system can not be used. The concentration level and cognitive effort required for each participant for the evaluation task has also to be taken into account. For these different reasons, the set of speech sequences has to be limited in size to make the task feasible and efficient while relevant for the assessment of the quality of the automatic system decisions. First, records presenting low signal quality, noise or other artifacts are excluded from the study. Then, on the basis of the automatic annotation, speech sequences are selected on the expertise of two human annotators (distinct from the jury used later). Each speech sequence contains from one to 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 annotators.

---

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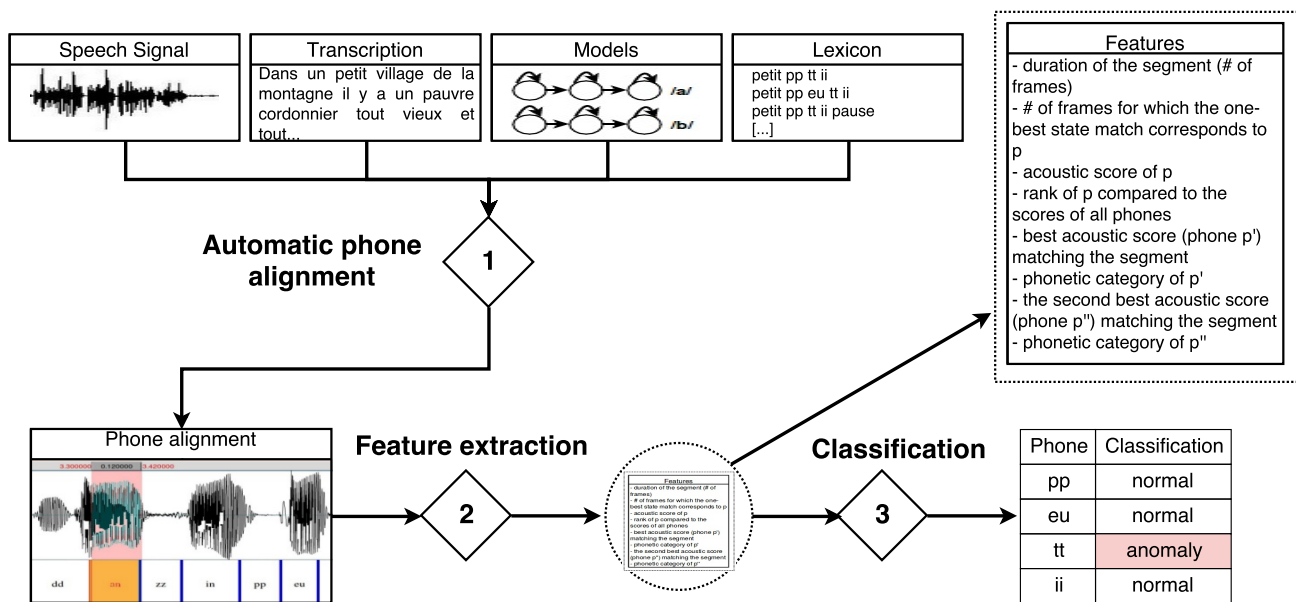
Figure 1: Automatic approach for a phone-level anomaly detection.

Speech sequences are chosen to fit one of the following 4 predefined categories, regarding uniquely the target word(s) (as noted above, the rest of the words in the speech sequences is considered as normal, independently of the categories) :

- 12.5% of the test sequences are referred to as "obvious segments". Here, both annotators agree 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% are referred to as "ambiguous segments". Here, the human annotators disagree with each other and are not able to decide whether the automatic annotation of the target word(s) as abnormal is correct or not;

- 25% are referred to as "false negatives". Here, both annotators consider that **the system fails to detect the presence of a true anomaly on the target word(s)**;

- 25% are referred to as "false positives". Here, both annotators consider that the system falsely labels the target word(s) as abnormal.

Other factors shape the set of the speech sequences. First of all, efforts have been concentrated on selecting speech produced by the largest number of patients, and representing the four pathologies available on our corpora. Second, efforts are 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, included 33 dysarthric patients and 7 healthy control speakers, are finally selected for the perceptual evaluation task.

The last feature of the protocol relies on the choice of the participant listeners and their degree of expertise to evaluate the presence or not of abnormal words in the speech sequences. Let's recall that this perceptual evaluation protocol aims at evaluating the quality of the outputs of an automatic system, considered itself as an "expert" - its goal is to bring some objective "expertise" to clinicians or phoneticians in their analysis of speech disorders. It seems natural to demand that listeners, qualified in evaluating such speech disorders, participate in this evaluation protocol. A jury of expert listeners are, therefore, selected.

3. Experimental protocol

3.1. Corpora

All the selected speech sequences are extracted from French read speech recordings of the fairy tale "*Tic Tac*" (The elves and the shoemaker). In total, 98 sequence produced by 40 speakers from dysarthric speech corpora *VML* and *TypALoc* (Meunier et al., 2016) are selected. Four pathologies are represented in these corpora : Parkinson's Disease (PD), Cerebellar Ataxia (CA), Amyotrophic Lateral Sclerosis (ALS), and lysosomal diseases (LYS).
Table 1 details the number of patients and sequences for each pathology and their dysarthric class. The selected segments were extracted from the recordings using Praat (Boersma and Weenink,) and artificial silences of 400 and 200ms were added to each at the beginning and the end respectively in order to avoid abrupt signal cuts for the perceptual evaluation process.

3.2. Jury and experimental design

The selected jury includes 29 experts aged between 22 and 58 (average age of 33). They all have French as their mother tongue and have no prior audition or learning disorders. The jury is composed of 10 speech therapists, 18 final-year speech therapy students and 1 Ear, Nose and Throat (ENT) specialist, and speech pathologist.

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.

The proposed perceptual evaluation task is then computerized using the Perceval platform (Ghio et al., 2003). The evaluations last between 25 and 40 minutes and are performed in quite rooms as follows: (1) participants are presented with an instruction list on the screen explaining the test procedure; (2) an oral instruction indicating to focus only on articulatory realization and not to take prosodic or vocal aspects into account is then given to all participants; (3) a training phase of 4 sequences is proposed in order to get participants familiarized with the task and the use of Perceval platform; (4) when the evaluation starts, an orthographic transcription of the sequence appears on the screen. Under each word, the expert has to check one of two boxes to label the word as "**normal**" or "**abnormal**". Figure 2 shows an example screen shot of the experiment. The sequences are presented in a totally randomized order for each participant and can be played up to 3 times before making an evaluation.
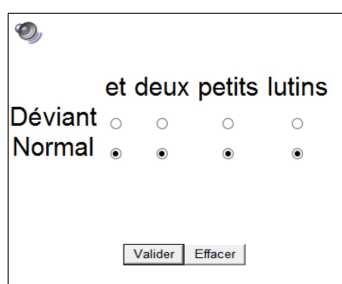

Figure 2: Screen shot from the Perceval platform used in the perceptual evaluation.

3.3. Evaluation methodology

To analyze the perceptual evaluation results, 3 System-Jury agreement rates are computed:

- $AG\_targetAnomaly$ rate, measures the System-Jury agreement rate on the targeted abnormal words (target words 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;

- $AG\_targetNormal$ rate, measures the System-Jury agreement rate on the targeted normal words (target words automatically labeled as normal) for the "false negatives" category. This rate reflects the system inability to detect potential present anomalies (according to the couple of 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;

- $AG\_nonTargetNormal$ rate, measures the System-Jury agreement rate on the non target words (automatically labeled as normal and considered as such during the sequence selection by both the annotators) for all test categories ("obvious segments", "ambiguous segments", "false positives" and "false negatives"). This rate will measure 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 and discussions

4.1. Results according to test sequence categories

Figure 3 depicts the distribution of the different agreement rates when computed on each test category.

Considering the $AG\_targetAnomaly$ measure, we observe a strong heterogeneity in the results depending on the test category, reaching 78%, 58%, and 13% for "obvious segments", "ambiguous segments" and "false positives" categories respectively. The high $AG\_targetAnomaly$ rate on "obvious segments" confirms the capacity of the automatic approach to detect highly distorted segments. This capacity has been already highlighted in (Laaridh et al., 2015a) with 81% of phone-based anomalies annotated by an expert well detected by the system.

In contrast, the low $AG\_targetAnomaly$ rate of 13% observed on "false positives" reveals the limitations 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 they could be related to acoustic distortions
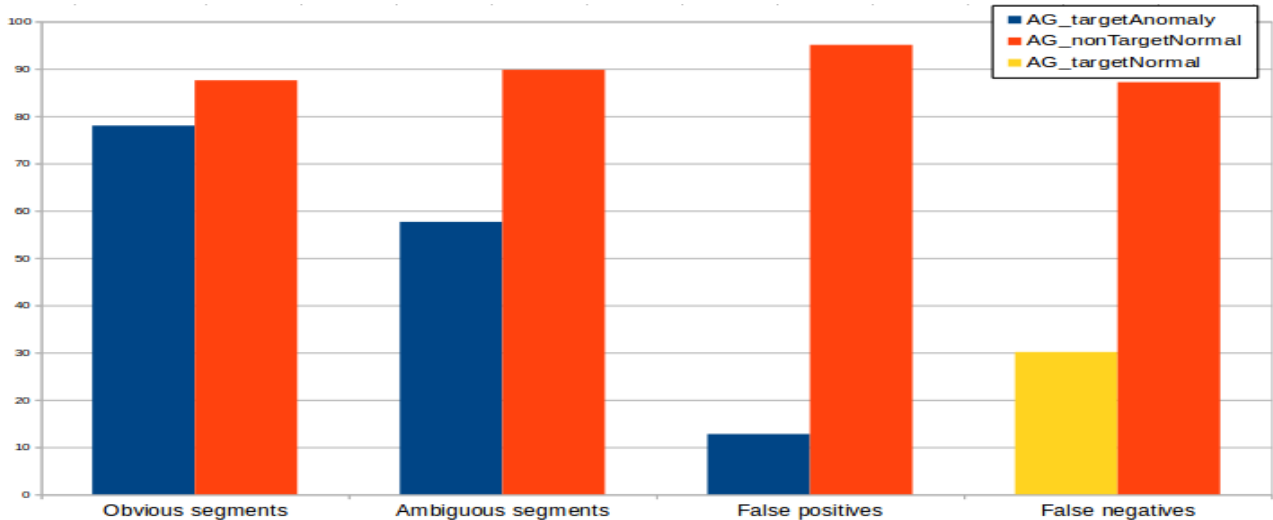
Figure 3: System-Jury agreement rates (%) per test sequence category.

| Population | Obvious segments | | Ambiguous segments | | False negatives | | False positives | |
	AG_target Anomaly	AG_nonTarget normal	AG_target Anomaly	AG_nonTarget normal	AG_target Normal	AG_nonTarget normal	AG_target Anomaly	AG_nonTarget normal
CTRL	81.0	99.1	15.2	99.7	50.6	97.5	1.4	100.0
CA	71.3	92.3	59.8	86.4	24.9	86.6	9.8	94.6
PD	78.2	89.1	42.7	93.7	64.4	91.2	8.6	97.3
ALS	74.6	52.9	79.0	77.1	25.6	75.3	19.6	98.0
LYS	98.3	81.9	68.1	86.6	8.6	72.9	15.5	85.6

Table 2: System-Jury agreement rates (%) on automatically detected abnormal ($AG\_targetAnomaly$) and normal ($AG\_targetNormal$ and $AG\_nonTargetnormal$) words per population and test sequence category.

of non-pathological nature (noise, breaths, etc.). Other hypotheses can be advanced to explain this behavior such as the presence of true anomalies not detected by the human experts in these segments or the presence of erroneous data in the system training corpora (errors in the human annotation used as reference).

Nonetheless, we support that this observation has to be balanced. Indeed, the System-Jury agreement rate on the non target normal words $AG\_nonTargetNormal$, aggregated over all the test categories, reaches 88% and confirms that the behavior of the automatic approach is far from being arbitrary. Therefore, the system behavior on the "false positives" category being finally restrained to a small amount of speech segments, and judged as marginal, it would be more appropriate to describe the behavior of the system as more severe than the human experts.

Considering the two remaining categories, "ambiguous segments" and "false negatives", the System-Jury agreement rates confirm the difficulty and the non trivial nature of the perceptual evaluation of dysarthric speech task even when performed by experts. Here, more than half (58%) and nearly third (30%) of the jury decisions agree with the system on both categories respectively, which demonstrates a strong heterogeneity in the decisions given by the experts themselves.

4.2. Inter-population variability

Table 2 details the System-Jury agreement rates per population and test category. We can note that the best $AG\_targetAnomaly$ rate is computed over LYS patients reaching 98.3% and 68.1% on "obvious segments" and "ambiguous segments" respectively. This behavior can be expected considering that this population is involved in the modeling of the abnormal phones in our system and is consistent with previous results in (Laaridh et al., 2015a). This does also highlight the importance of the training phase in our automatic approach and suggests that the use of more data associated with different pathologies and dysarthric classes should improve the system performance, already very promising given the results reported earlier.

Considering the other populations, notable differences are observed between the different pathologies. This is highly important considering that the instructions given to the jury explicitly restrict the evaluation task to the articulatory production of speakers. This behavior is particularly evident on ALS patients on whom the jury annotated the most anomalies compared to other populations and where the $AG\_targetAnomaly$ rate reaches 19.6% on the "false positives" category. In contrast, an opposite behavior is observed on CTRL speakers and PD patients for whom an overall good quality of the speech is usually observed and the computed $AG\_targetAnomaly$ rate over the "ambiguous segments" reaches 15.2% and 42.7% respectively.
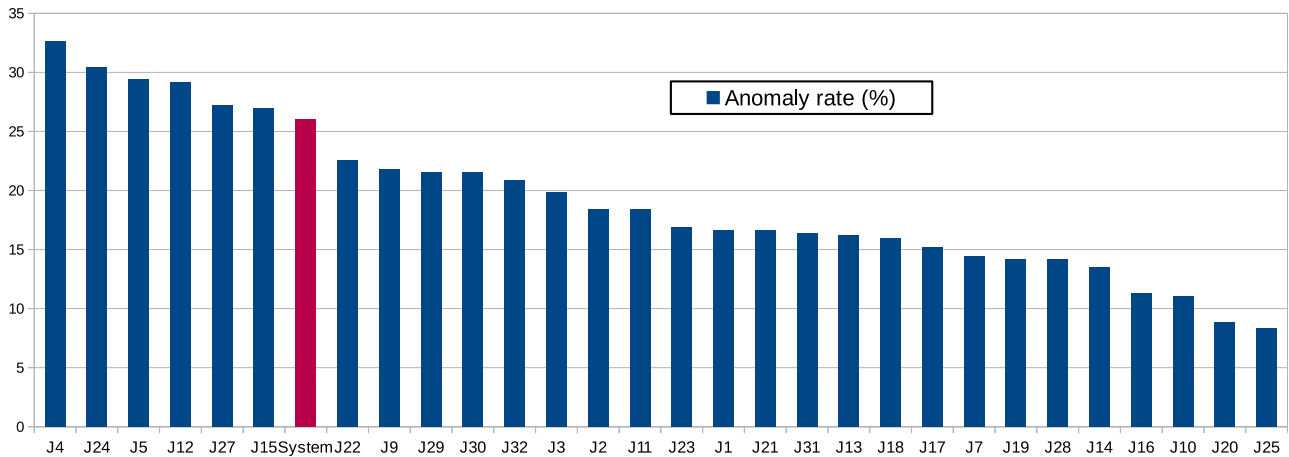
Figure 4: Anomaly rate (%) per jury member (perceptual, blue bars) and for the system (red bar).

4.3. Jury responses

Even though the experimental protocol proposed here aim at studying the behavior of an automatic anomaly detection approach, the computed results highlighted the difficulty for a human, even expert, to apprehend and detect deviations at the word level in dysarthric speech and revealed some evaluation tendencies depending on patients' pathologies and dysarthria severity. Figure 4 depicts the perceptual anomaly rate (%) per jury member.

This tends to show that listeners may be influenced by the contextual information. If speech sounds pathological (ALS patients for instance) then anomalies are more often detected in words. We should therefore ask if listeners are able to perform the same task as the system; the system is able to focus on short units to detect anomalies (phones, syllables, words) while subjects perform a contextual task to take their decision. The 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. And this is also not the way we perceive speech. The process of speech perception requires a large context of speech in order to evaluate if it is distorted or not. The need to focus on a specific item is a very difficult task for listeners.

5. Conclusions

This paper investigates the results of a perceptual evaluation of the annotation performed by an automatic anomaly detection system on dysarthric speech. The results confirm the capacity and relevance of the automatic approach in detecting the presence of anomalies in dysarthric speech (high $AG\_targetAnomaly$ rates on "obvious segments"). Moreover, and even on the more nuanced anomalies ("ambiguous segments"), the jury agrees 58% of the time with the automatic approach decisions.

However, the low $AG\_targetAnomaly$ rate computed over the "false positives" category confirms the approach tendency to be more severe than the human experts. Consider-

ing the limits of the perceptual evaluation recognized in the literature(Zyski and Weisiger, 1987; Fex, 1992), we suggest that a more primitive question must be raised: should an automatic approach replicate the behavior of a human expert and what place should be envisaged for future investigations between supervised (relying on human annotations) and semi- or unsupervised approach ?

6. Acknowledgments

7. Bibliographical References

Auzou, P. and Rolland-Monnoury, V. (2006). *Batterie d'évaluation clinique de la dysarthrie*. Édition Ortho.

Boersma, P. and Weenink, D.). Praat: doing phonetics by computer. http://www.praat.org/.

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

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

Darley, F. L., Aronson, A. E., and Brown, J. R. (1975). *Motor speech disorders*. W. B. Saunders and Co., Philadelphia.

Duffy, J. R. (2005). *Motor speech disorders: substrates, differential diagnosis and management*. Motsby- Yearbook, St Louis, 2nd edition.

Fex, S. (1992). Perceptual evaluation. *Journal of voice*, 6(2):155–158.

Ghio, A., André, C., Teston, B., and Cavé, C. (2003). 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)*, 22:115–133.

Laaridh, I., Fredouille, C., and Meunier, C. (2015a). Automatic detection of phone-based anomalies in dysarthric speech. *ACM Transactions on accessible computing*, 6(3):9:1–9:24, May.

Laaridh, I., Fredouille, C., and Meunier, C. (2015b). Automatic speech processing for dysarthria: A study of inter-pathology variability. In *Proceedings of Intl Congress of Phonetic Sciences (ICPHs'15)*, Glasgow.

Laaridh, I., Fredouille, C., and Meunier, C. (2016). 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)*, Portoroz, Slovenia, may.

Meunier, C., Fougeron, C., Fredouille, C., Bigi, B., Crevier-Buchman, L., Delais-Roussarie, E., Georgeton, L., Ghio, A., Laaridh, I., Legou, T., Pillot-Loiseau, C., and Pouchoulin, G. (2016). 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)*, Portoroz, Slovenia, May.

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

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Zyski, B. J. and Weisiger, B. E. (1987). Identification of dysarthria types based on perceptual analysis. *Journal of Communication Disorders*, 20(5):367–378.