Automatic Detection of Phone-Based Anomalies in Dysarthric Speech

Imed Laaridh, Corinne Fredouille, Christine Meunier

A

Automatic detection of phoneme-based anomalies in dysarthric speech

IMED LAARIDH, University of Avignon, CERI/LIA; University of Aix Marseille
CORINNE FREDOUILLE, University of Avignon, CERI/LIA
CHRISTINE MEUNIER, University of Aix Marseille, CNRS, LPL UMR 7309, 13100, Aix-en-Provence, France

Perceptual evaluation of speech is still the most used method in the examination and the longitudinal evaluation of patients suffering from speech disorders in clinical practice. Even abundant research work has been done on acoustic analysis of speech productions exhibiting impairments in order to bring and enhance knowledge in regards to the acoustic alterations which can be observed, more descriptive analysis is required to take into account their large variability, considering patients suffering from the same disease or across different diseases.
In this context, this paper proposes to investigate automatic speech processing approaches, dedicated to the detection and localization of abnormal acoustic phenomena in speech signal produced by patients suffering from speech disorders. This automatic process aims at limiting and enhancing the manual investigation of some human experts while scrutinizing speech signal by focusing their attention on specific parts of speech, considered as atypical from an acoustical point of view.
The experimental evaluation of two different approaches for the task of detecting acoustic anomalies, conducted on two different corpora comprising both dysarthric and control speakers, demonstrates very promising results and the potentiality of approaches to be applicable to different types of dysarthria and neurological diseases.

General Terms: Speech disorders, dysarthria, automatic speech processing, objective detection of acoustic anomalies, supervised classification

ACM Reference Format:
Imed Laaridh, Corinne Fredouille and Christine Meunier, 2014. Automatic detection of phoneme-based anomalies in dysarthric speech. *ACM Trans. Access. Comput.* V, N, Article A (January YYYY), 16 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Dysarthria is a motor speech disorder, consequence of some neurological damages located either in the central or in the peripheral nervous system. This may result in disturbances in any of the components involved in speech production like respiratory, phonatory, resonatory, articulatory and prosodic. Consequently, this may reflect weakness, spasticity, incoordination, involuntary movements, or variable, from excessive to reduced, muscle tone [Darley et al. 1969; Murdoch 1998; Duffy 2005], depending on the neurological damage localization. Research on dysarthria is very abundant, covering different domains and aspects : perceptual evaluation ([Enderby 1983][Murdoch 1998][Hustad 2008][Lowit and Kent 2010]), acoustic analysis ([Kent et al. 1999][Rosen and R. D. Kent 2006][Christina et al. 2012]), automatic intelligibility assessment ([Middag et al. 2009][MyungJong and Hoirin 2012]), automatic speech recognition ([Rosen and Yampolsky 2000][Strik et al. 2002][Parker et al. 2006][Christensen et al. 2013]).
Even if a limited set of typical acoustic-perceptual cues including, for instance, imprecise consonants, vowel centralization, slow rate, monopitch, monoloudness, hypernasality, is commonly accepted to characterize the main disturbances of the various types of dysarthria in speech production, more descriptive acoustic and phonetic analysis is still necessary in order to take into account the large variability in terms of speech alterations observed among patients in different disease groups but also inside the same group [Tomik and Guiloff 2010]. Indeed, in addition to the localization of the neurological damages and the associated disease, speech production alterations (and

acoustic consequences) may be influenced by the onset range of symptoms (patients with Parkinson's disease may not suffer from dysarthria at all even other symptoms have appeared. In a same way, as reported in [Tomik and Guiloff 2010], the time between the onset of speech symptoms and the diagnosis in patients suffering from amyotrophic lateral sclerosis may range from 33 months prior the diagnosis to 66 months after the diagnosis [Yorkston et al. 1993]), by the progression of disease, which can be very patient-dependent, or by the patient himself/herself, considering the compensation/adaptation strategies he/she develops to overcome speech impairment. On the other side, more and more recordings of patients are available for research purposes [Fougeron et al. 2010][Ghio et al. 2012], some of them recently distributed [Rudzicz et al. 2012], and interest for more contrasting speech material like spontaneous speech (as opposed to sustained vowels or word repetition) for instance is growing.

Research work presented here contributes to a larger research project dedicated to the automatic detection and localization of abnormal acoustic phenomena in speech signal produced by patients suffering from different types of dysarthrias. This automatic detection and localization aims at limiting and enhancing the manual investigation of some human experts while scrutinizing speech signal. Indeed, this automatic process should permit to treat a larger amount of speech production while focusing human experts on specific parts of speech, considered as atypical. This process is notably interesting for speech production of patients suffering from mild to moderate dysarthria for which speech impairment may be scattered across speech signal. Moreover, this automatic detection and localization of abnormal acoustic phenomena can have some applications in clinical practice. Indeed, the evaluation of dysarthria by clinicians could be partially helped by a visual display of abnormal phenomena localized in speech signal of patients, like a map. In a same way, maps should be relevant to compare speech productions of a patient in time, during clinical treatment or rehabilitation for instance. Finally, this automatic process could be extended to other kinds of speech disorders implying acoustic alterations in speech signal, like larynx or head cancers for instance.

In [Chandola et al. 2007], it is reported that anomaly detection refers to the problem of findings patterns in data that do not conform to expected behavior. Applications for anomaly detection are numerous like intrusion detection, fraud detection, medical condition monitoring, fault diagnosis in instrumentation, ... Still in this review paper, the authors expose the challenging task of detecting anomalies through different factors, which, for some of them, can be easily connected to speech analysis as follows : (1) the boundary between normal and abnormal behavior is often not precise, leading to false positive detection or missed, (2) when anomalies come from malicious action (like a fraud), adaptation strategies can be employed in order to mask anomalies (related to compensation strategies used by some patients to overcome their speech impairment), (3) labeled data (including anomalies), necessary to train some models used by anomaly detection systems and to evaluate them, are not available or in a very sparse way, (4) data may contain noise, which can be taken for anomalies and, therefore, difficult to distinguish and remove. In the field of anomaly detection, two main groups of techniques can be identified. The first group refers to techniques which model the normal behavior uniquely. When an incoming data has to be processed, the system has to determine whether it may be covered by the normal class model. The second group refers to techniques which model both the normal and abnormal behaviors. These techniques require labeled data in order to train both models. Then, the system has to determine which model an incoming data belongs to.

In a previous work [Fredouille and Pouchoulin 2011], the authors proposed a baseline system dedicated to the detection of abnormal zones in dysarthric speech. The proposed method can be related to the first group of techniques reported above for the

anomaly detection (modeling of normal data uniquely). In this paper, this work is extended by proposing an approach related to the second group of techniques (modeling of normal and abnormal data) and by comparing it to the former. As labeled data for dysarthric speech is rare and very time-consuming, the experimental validation of the novel system will also examine the relationship between the anomaly detection decisions made by the automatic system and results of a perceptual assessment made by a jury of experts on additional speech productions issued from dysarthric patients, for which no labeled data is available.
The rest of this paper is organized as follows. In section 2, both automatic approaches for the anomaly detection task are described. The evaluation protocol as well as the two corpora involved in this paper are presented in section 3. In section 4, proposed approaches are compared in terms of detection performance on labeled data. The behavior of the novel approach is then discussed in regards to the perceptual assessment of additional dysarthric speech production. Section 5 provides a conclusion and some direction for future work.

2. AUTOMATIC DETECTION OF PHONEME-BASED ANOMALIES

This section describes the automatic approaches proposed for the task of acoustical anomaly detection in speech signal. While the first approach is focused on the modeling of information related to normal speech only, the second approach deals with information extracted from both normal and abnormal parts of speech. Due to its simplicity, the first system is considered as a baseline in the rest of the paper.

In this paper, the authors focus the anomaly detection on the phoneme level for two main reasons : (1) the phoneme duration is considered as sufficient to provide usable information, notably compared with smaller units like the frame unit typically used in speech processing domain, (2) phonemes could be acoustically distorted due to voice alteration, movement disorders and/or articulatory impairments, typical in dysarthric speech. Consequently, either for a test phase or training and test phases, both approaches will share a common step involving an automatic text-constrained phoneme alignment, described below.

2.1. Text-constrained phoneme alignment

The segmentation of speech utterances in phonemes is carried out thanks to an automatic text-constrained phoneme alignment tool. This one takes as inputs the sequence of words pronounced in a speech utterance, a phonetized lexicon of words coupled with different phonological variants, based on a set of 37 French phonemes. The automatic speech processing is then based on a Viterbi decoding and graph-search algorithms, which the core is the acoustic modeling of each phoneme, based on Hidden Markov Models (HMM) (see [Brugnara et al. 1993] for more details). In this paper, a 3-state context-independent HMM topology is used for each phoneme model, estimated from French radiophonic speech recordings [Galliano et al. 2005]. The sequence of words is issued from an orthographic transcription performed by human listeners following some specific annotation rules able to take into account deletions, substitutions and insertions of word and/or sequences of phonemes.
The speech segmentation results in a couple of start and end boundaries per phoneme present in produced speech signal.

2.2. Baseline system : normal speech modeling

The methodology proposed in this section relies on three main steps. The first one consists in segmenting the speech signal in homogeneous zones from which detection decision is made. The second step permits to label each of these homogeneous zones as abnormal or not. The third one provides a visual display to make the ”reading” of the

abnormal speech zones, provided by the automatic detection, simpler and easier. Next sections will provide details for the two latter steps, the first segmentation step, relying on the text-constrained phoneme alignment process, being described previously.

2.2.1. Acoustic score measurement. The goal of this second step is to associate each phoneme with a normalized acoustic score, further used to determine a normality degree. In this paper, this normalized acoustic score is defined as follows:

$$LL_p^{norm}(y_p) = log(\frac{L_p^{Constrained}(y_p)}{L^{Unconstrained}(y_p)}) \quad (1)$$

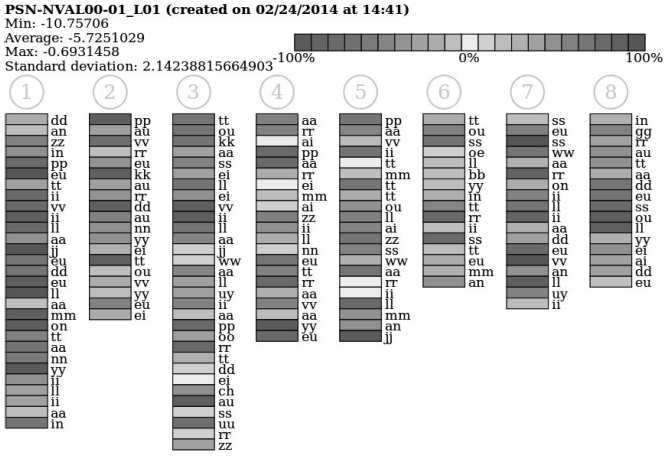
where $LL_p^{norm}(y_p)$ is the expected normalized acoustic score computed for given phoneme p on related speech segment y_p. $L_p^{Constrained}(y_p)$ is an acoustic score assigned to phoneme p by the automatic text-constrained phoneme alignment process. More precisely, this score is the likelihood computed from the speech segment y_p and the HMM model of phoneme p. In order to be able to compare acoustic scores given any phoneme, $L_p^{Constrained}(y_p)$ is normalized by dividing it by $L^{Unconstrained}(y_p)$. The latter is also an acoustic score, especially designed for score normalization. It represents the best HMM state sequence that the Viterbi algorithm might produce, considering no constraint on text pronounced nor temporal constraints between HMM states (referred to as text-unconstrained alignment process). In this context, $L^{Unconstrained}(y_p)$ can be defined as:

$$L^{Unconstrained}(y_p) = \sum_{t=1}^{T_{y_p}} \arg\max_{k \in K} L_k(y_p^t) \quad (2)$$

with K the set of HMM states available, T_{y_p} the number of frames in speech segment y_p, and $L_k(y_p^t)$ the likelihood computed between the t^{th} frame issued from y_p and the k^{th} HMM state.
The closer $L_p^{Constrained}(y_p)$ and $L^{Unconstrained}(y_p)$ are, the closer to 0 the normalized score ($LL_p^{norm}(y_p)$) is. In this case, both automatic text-constrained and text-unconstrained alignment processes converge towards the same phoneme p. Conversely, the more distinct both scores are, the closer to $-\infty$ normalized score is. Here, both automatic text-constrained and text-unconstrained alignment processes converge towards different phonemes p and p'.

2.2.2. Normality scores and cartography. The last step of the methodology exploits the normalized acoustic scores assigned to each phoneme by determining whether a phoneme has to be considered as normal or abnormal. This decision is made from the speech production point of view, typically in the context of speech disorders. It is based on the computation of a degree of normality per phoneme as follows. Firstly, considering a set of healthy speakers and their speech utterances, a normalized acoustic score per phoneme pronounced is computed according to steps 1 and 2. Secondly, considering that this set of scores represents normal phonemes (to oppose to abnormal ones), descriptive values like minimum, maximum, mean figures are computed to design a very straightforward reference scale of normality. Thirdly, considering the speech utterance of a patient suffering from speech disorders, the set of phoneme-based normalized acoustic scores associated with is projected onto the reference scale of normality. This results in a score of normality per phoneme, ranged in [-100;100] interval. Finally, the set of scores of normality computed for a speech utterance can be easily reported graphically on a normality cartography to facilitate its observation, as illustrated in figure 1. On this figure, the reference scale is reported on the top-right through a blue to red color gradation: the blue color representing normality and the

Fig. 1. Example of cartography relating to a male patient, illustrating the degree of normality per phoneme



red one the presence of anomalies, the more the color tends towards red, the more abnormal the phoneme. Each column represents a sequence of phonemes - themselves depicted as individual rectangles - produced in a "pseudo-sentence"; the set of columns represents the entire text read by the speaker.
This graphical representation permits in a simple way to determine in which zones of the speech signal the automatic processing has detected anomalies and to know which phonemes are associated with.

2.3. Normal and abnormal speech modeling

The second approach presented here for the phoneme-based anomaly detection aims at taking both normal and abnormal information into account and at discriminating them. As detailed below, it also relies on three steps, integrating firstly, a text-constrained phoneme alignment, similarly to the baseline system. The second step aims at characterizing each phoneme with a set of features, considered as relevant for the discrimination task. The final step consists in a supervised classification task based on two classes - normal and abnormal phonemes - which will permit to label as normal or abnormal (detection of an anomaly in the latter) each incoming phoneme issued from an unseen speech signal.

2.3.1. Feature Extraction. The detection process being focused on phonemes, the set of features designed for the discrimination task between normal and abnormal phonemes is mainly derived from the automatic phoneme alignment outputs and the 37 HMM models associated with. Indeed, for each phoneme p and the associated speech segment s_p (defined by its start and end boundaries), issued from the automatic text-constrained alignment, following features are extracted :

— the phoneme duration, expressed in terms of number of 10ms frames present in s_p,
— the number of frames in s_p for which the one-best state search among the HMM-based phoneme models, applied at the frame level, corresponds to those of p,
— the acoustic score of the one-best phoneme p', while comparing scores of all the HMM phoneme models computed on segment s_p. If p is the one-best phoneme, the second-best is considered,
— the phonetic category of p',

— the acoustic score of the second-best phoneme p'', while comparing scores of all the HMM phoneme models computed on segment s_p. If p is one of the two best phonemes, the third one is considered,
— the phonetic category of p'' ,
— the rank of p and its acoustic score while comparing scores of all the HMM phoneme models.

All of these features are used to characterize each phoneme during the training and the classification phases. For this reason, all the acoustic scores are normalized.

2.3.2. SVM-based supervised classification. SVM learning theory has been largely applied on pattern recognition problems. It is based on the search of the surface which will best separate a set of data in different classes so that the margins between them are maximal. This surface is known as the optimal hyperplane [Vapnik 1995][Scholkopf and Smola 2001].
Here, SVM method is applied on a two-class problem, discriminating data related to normal phonemes and those to abnormal phonemes (anomalies). Abnormal labels associated with phonemes are issued from a manual annotation carried out by a human expert. Features defined previously are used as incoming data for each phoneme.
Unlike the baseline system, different SVM sub-systems are trained here by distinguishing the male speech productions from the female ones. Moreover, specific attention is given to four different phonetic categories - unvoiced consonants, voiced consonants, oral vowels, nasal vowels - for which separate SVM sub-systems are trained. This is motivated by the investigation of finer abnormal and normal phenomena models and the acoustic specificities of each phonetic category. On the other side, each SVM sub-system is trained with a more limited amount of labeled data.

In this way, each phonetic category has its own phoneme set, issued either for the female speech productions or the male ones, used to train its proper SVM model. This model will be used only to classify this category of phonemes during test. Due to the limited size of labeled data available, notably regarding abnormal phonemes (compared with normal ones), a balanced number of samples representing both classes is given to the SVM process to estimate models. Therefore, averaged values of about 750, 680 and 420 normal and abnormal phonemes are used to train models of voiced consonants, oral vowels and unvoiced consonants respectively. As nasal vowels are really sub-represented in the text read by the speakers, an averaged value of about 90 phonemes only are used to train models of nasal vowels. Besides, the ”leave-one-out” technique is applied, which consists in a circular evaluation : each time a speech recording is considered as a test sample, it has to be excluded from the training set used to estimate class models. This technique permits to train classification models on larger data sets and, at the same time, to exploit all speech recordings annotated by the human expert in the evaluation procedure of the detection system.
The different SVM sub-systems are implemented thanks to the SVMlight tool (see [Joachims 1999] for more information).

3. EXPERIMENTAL PROTOCOL

This section describes the experimental protocol used to evaluate the behavior of approaches, described previously, for the task of detection of anomalies in speech signal. This experimental protocol comprises two different databases used in this study.

3.1. Databases

3.1.1. Corpus 1. The current study is based on a speech corpus recorded at the hospital La Pitié-Salpétrière in Paris. 8 dysarthric patients (4 women and 4 men), aged

Table I. Information related to the patients, including the number of recordings, the number of phonemes pronounced as well as the number of phonemes annotated as abnormal by the human expert. Figures are averaged over the different recordings available per patient..

Patients	Recording Nb	Averaged nb. of phonemes	Averaged nb. of abnormal phonemes	% abnormal phonemes
Male 1	4	532	50	9,5
Male 2	3	116	43	37,5
Male 3	5	549	84	15,3
Male 4	6	268	80	30.2
Female 1	5	529	87	16,5
Female 2	5	100	77	76,3
Female 3	4	540	82	15,2
Female 4	3	306	102	33,5

from 23 to 43 years, and 6 control subjects participate to this corpus. Patients suffer from rare lysosomal storage diseases and show disparities in the severity degree of dysarthria in regards to the progression of their disease.
Patients were recorded for 2 years with a recording each six month approximately. The control subjects were recorded for 1 month with a recording each week. This schedule results in 3 to 6 recording sessions per speaker. Participants were asked to read the same text, a French fairytale called “Le cordonnier” (The cobbler), as naturally as possible. The duration of speech utterances varies from 48s to 196s, with an average of around 60s for control subjects, and 85s for patients. Differences observed in the patient durations are due to differences in rhythm as well as the degree of fatigability of certain patients, which prevents them to read the entire text. The read text contains about 550 phonemes (including about 290 consonants and 260 vowels).
All the speech utterances related to the patients were analyzed by a human expert in order to annotate the abnormal speech zones at the phoneme level. Helped with the listening and the Praat-based analysis of the speech signal coupled with the automatic phoneme segmentation (see section 2.1 for details), the main task of the expert was to label a phoneme as normal or abnormal, by indicating in this last case, the type of alteration (noise, voicing impairment, spectral distortion, ...). In the same time, it was requested from him to make some frontier corrections in the phoneme segmentation when required. Table I provides detailed information per patient (figures are averaged over the different patients’ recordings available) relating to the number of recordings, of phonemes and of abnormal speech zones annotated by the expert.

3.1.2. Corpus 2. In addition to the initial speech corpus, one has employed a second database containing 118 dysarthric and normal speakers (53 women and 65 men) aged from 32 to 76 years gathered within the DesPhoAPaDy project [Fougeron et al. 2010]. Unlike the first corpus where patients suffered only from rare lysosomal storage diseases, this second database presents various diseases and severity degrees of dysarthria:

— 37 patients (23 women and 14 men) with Amyotrophic Lateral Sclerosis (ALS),
— 31 patients (8 women and 23 men) with Parkinson’s disease.
— 21 patients (8 women and 13 men) with cerebellar ataxia.

In addition, 29 healthy people (14 women and 15 men) are considered as control speakers.

Patients and control speakers were recorded within soundproof rooms during a phoniatry consultation, across different medical institutes. All speakers were recorded reading the same text as corpus 1: ”Le cordonnier” story. Table II provides detailed

Table II. Information related to speakers of corpus 2, including the averaged values of global severity degree of dysarthria, of articulation impairment rate, of intelligibility measure and of speech rate, issued from the 11 experts' perceptual evaluation,.

Disease	Number	Averaged nb. of phonemes	Global dysarthria severity degree	Articulation impairment	Intelligibility	Speech rate	
						Slow	Fast
Female Parkinson's disease	8	571	0.65	0.49	0.43	0	0.81
Female ALS	23	538	1.91	1.66	1.25	-1.46	0
Female Cerebellar ataxia	8	586	1.51	0.50	1	-1.27	0.76
Female lysosomal storage disease	4	377	2.2	1.9	1.6	-1.4	2.8
Female control speaker	14	566	0.10	0.03	0.05	-0.30	0.50
Male Parkinson's disease	23	603	1.04	0.85	0.75	-0.95	1.09
Male ALS	14	548	1.86	1.68	1.30	-1.39	0.82
Male Cerebellar ataxia	13	597	1.46	1.33	0.93	-1.31	0.18
Male lysosomal storage disease	4	374	1.9	1.7	1.5	-2.1	0.8
Male control speaker	15	561	0.16	0.12	0.03	-0.25	0.31

information about this corpus.

Unlike corpus 1, no human annotation of abnormal phonemes was carried out in corpus 2. Recordings were instead evaluated perceptually by a jury of 11 experts (7 to 26 years of experience in dysarthric speech perceptual evaluation). Experts were asked to rate all the speakers (including control speakers) on a large set of perceptual items, including the following ones this paper focuses on (the reader may refer to [Lhoussaine 2012] for a complete list of perceptual items, rated by the experts):

— global evaluation of the dysarthria severity degree rated on a scale from 0 to 3 (0 - no dysarthria to 3 - severe dysarthria),
— evaluation of speech intelligibility on a scale from 0 to 3 (0 - good intelligibility to 3 - unintelligibility),
— evaluation of articulation impairment on a scale from 0 to 3 (0 - normal to 3 - pronounced and constant articulation alterations),
— evaluation of speech rate on a scale from -3 to 3 (-3 - very slow, 0 - normal, to 3 - extremely fast speech rate).

Finally, it is interesting to note that corpus 1 was also partially evaluated by this expert jury in the same assessment schedule, and, therefore, under the same conditions. This concerns all the dysarthric patients, evaluated on their first speech recording.

3.2. Evaluation Protocol

3.2.1. Measurements. The reliability of approaches proposed for the detection of abnormal phonemes (anomalies) is evaluated by comparing their detection outputs with the human expertise, which remains the Gold standard for this tricky, but subjective task. Based on the expert's annotation, this comparison permits to compute different evaluation measures between (normal and abnormal) labels coming from the automatic approaches and the human expert.
We propose two main measures, stemming from the retrieval information domain [Makhoul et al. 1999], focused on the detection of the abnormal phonemes :

— the abnormal class-based recall measure, ranged from 0 to 1, named $AbnRecall$, given by the ratio between the number of zones well detected as anomalies by the automatic processing and the number of zones labeled as abnormal by the human expert. This ratio will measure the performance of the automatic processing in detecting abnormal zones correctly. The more close to 1 the ratio, the more the automatic system performs well to detect real abnormal zones.
— the abnormal class-based precision measure, ranged from 0 to 1, named $AbnPrec$, given by the ratio between the number of phonemes well detected as abnormal by the

automatic processing and the number of phonemes that the automatic processing labels as abnormal (truly or falsely). This ratio will measure the inverse rate of false alarm/false positive produced by the automatic methodology : the more close to 1 the ratio, the more precise in detecting the abnormal zones the automatic system .

It is worth noting that these measures focus on the detection of abnormal phonemes only. Similar measures could be used for the detection of the normal phonemes. Nevertheless, the imbalance in the number of normal and abnormal speech zones in the corpus used in this paper makes the latter meaningless (see table I). Second, $AbnRecall$ and $AbnPrec$ measures have to be considered as complementary in the assessment of the automatic detection processing. Indeed, if the recall score is necessary to measure the efficiency of the automatic approach in detecting abnormal speech zones, the precision score is important to measure the usability level of the automatic system in clinical or phonetics applications. Indeed, by assuming that the automatic detection system is tuned to label all the speech zones as abnormal, its recall measure should be equal to 1[1], giving it a perfect detection rate of abnormal speech zones. Conversely, its precision measure should be quite low (value depending on the total number of speech zones to process), which, in practical application, will focus the analysis on inappropriate speech zones.

The comparison between automatic and manual annotations is carried out following two strategies :

— strategy 1: comparing the annotations provided for each phoneme individually, without considering the local context : in this case we consider that we have a good match between the expert and the automatic system only if both automatic and expert labels matches on the same phoneme.
— strategy 2: comparing the annotations provided for each phoneme as well as previous and next phonemes (local context). In this case, if the human expert considers a given phoneme as abnormal while the automatic methodology detects an anomaly on the previous or next phoneme, but not on the given phoneme, then a right match is notified as well. This approach aims to support a ”one phoneme-based delay” in the automatic detection due to short boundary shifts in the automatic phoneme segmentation for instance.

Finally, the accuracy of both the automatic approaches while applied on control speakers or on corpus 2 cannot be evaluated via $AbnRecall$ and $AbnPrec$ measures due to the absence of expert's annotations on a phoneme level to compare with. Therefore, a simple agreement rate, based on the ratio between the number of speech zones labeled as normal by the automatic processing and the total number of phonemes will be used on the set of control speakers of corpus 1. This assumes that no abnormality is present in the control speakers' speech productions. In the case of speakers issued from corpus 2, the reliability of automatic approaches will be evaluated depending on the relationship, in terms of correlation rate, between the percentage of phonemes labeled as abnormal by the latter and the averaged rates issued from the perceptual evaluation carried out by the expert jury.

[1]if, at least one abnormal speech zone is present in the speech signal

Table III. Performance of baseline system applied on corpus 1 (dysarthric patients only), expressed in terms of abnormal class-based recall ($AbnRecall$) and precision measures ($AbnPrec$), according to comparison strategies 1 and 2.

	Strategy 1		Strategy 2	
Patients	$AbnRecall$	$AbnPrec$	$AbnRecall$	$AbnPrec$
Male 1	0.16	0.12	0.36	0.30
Male 2	0.44	0.38	0.77	0.77
Male 3	0.48	0.28	0.76	0.53
Male 4	0.44	0.37	0.73	0.64
Men average	0.38	0.29	0.65	0.56
Female 1	0.48	0.22	0.79	0.45
Female 2	0.43	0.66	0.87	0.98
Female 3	0.50	0.31	0.79	0.55
Female 4	0.60	0.36	0.88	0.63
Women average	0.50	0.39	0.83	0.65
Average	0.44	0.34	0.74	0.61

4. RESULTS

Both the baseline and SVM-classification methodologies presented in this paper have been applied on one or both databases presented in section 3.1. This section details the different results observed on these speech recordings.

4.1. Corpus 1

The baseline anomaly detection system described in 2.2 was applied on speech utterances of both dysarthric and control speakers of corpus 1. This results in a set of phoneme-based normalized acoustic scores and their corresponding values on the reference scale. The comparison with the human expert's annotations carried on a phoneme level was realized following the two mentioned comparison strategies. This permits to compute the performance evaluation measures proposed in section 3.2.1 and reported in table III. Measures given per speaker correspond to the average of values obtained on the various longitudinal recordings.

Observing $AbnRecall$ and $AbnPrec$ measures, the baseline system obtains averaged values of 0.5 and 0.39 for women patients respectively and 0.38 and 0.29 for men patients considering the first comparison strategy, and averaged values of 0.83 and 0.65 for women respectively and 0.65 and 0.56 for men patients according to the second phoneme-context strategy. These first results point out that the baseline approach reaches promising measures considering the second strategy even for men where the unique patient "male 1" contributes to considerably decrease averaged measure values.

The second automatic approach based on the supervised SVM-classification detailed on section 2.3 were also applied on the same recordings of corpus 1, respecting the leave-one-out principle.
Table IV provides the different measures reached according to both strict and one-phoneme-context comparison strategies.
Comparing results reached by both automatic approaches, we can observe that the SVM-classification based system outperforms the baseline : $AbnRecall$ measures have reached 0.89 and 0.72 averaged values on women and men patients respectively, exhibiting a 6% and 7% absolute improvement each. On the other hand, $AbnPrec$ presents a 5% absolute improvement for men patients reaching 0.61 while women $AbnPrec$ stagnates on 0.65 value. Similar behavior can be observed on the agreement rates reached by both approaches focusing on control speakers of corpus 1, as depicted in table V. Indeed, some improvements on the male speakers can be observed on the

Table IV. Performance of the SVM-Classification methodology applied on corpus 1 (dysarthric patients only), expressed in terms of abnormal class-based recall ($AbnRecall$) and precision measures ($AbnPrec$), according to comparison strategies 1 and 2.

	Strategy 1		Strategy 2	
Patients	$AbnRecall$	$AbnPrec$	$AbnRecall$	$AbnPrec$
Male 1	0.15	0.23	0.37	0.52
Male 2	0.47	0.24	0.90	0.68
Male 3	0.43	0.32	0.72	0.57
Male 4	0.54	0.33	0.89	0.65
Men average	0.40	0.28	0.72	0.61
Female 1	0.59	0.23	0.85	0.42
Female 2	0.83	0.68	1.00	0.98
Female 3	0.44	0.29	0.80	0.58
Female 4	0.60	0.34	0.90	0.60
Women average	0.62	0.39	0.89	0.65
Average	0.51	0.34	0.81	0.63

Table V. Performance of the baseline and SVM-based Classification approaches applied on the control speakers issued from corpus 1, expressed in terms of agreement rate measures (%).

Patients	Baseline system	SVM-based classification system
Control men average	92.8	97.6
Control women average	91.0	90.9
Control speaker average	91.6	93.1

Table VI. Performance of the SVM-Classification methodology applied on the corpus 1 (dysarthric patients only), expressed in terms of abnormal class-based recall ($AbnRecall$) and precision measures, ($AbnPrec$) computed on the different phonetic categories used for SVM sub-system training.

	Strategy 1		Strategy 2	
Phonetic category	$AbnRecall$	$AbnPrec$	$AbnRecall$	$AbnPrec$
Unvoiced consonant	0.68	0.41	0.86	0.61
Voiced consonant	0.60	0.31	0.84	0.57
Oral vowels	0.42	0.44	0.83	0.75
Nasal vowels	0.35	0.47	0.77	0.77

SVM-based system while results on the female speakers remain constant over both approaches.

Finally, table VI details $AbnRecall$ and $AbnPrec$ measures reached by the SVM-based classification approach, examining each phonetic category associated with individual SVM sub-system individually. Comparing these results, it can observed that the automatic system presents more difficulties in detecting anomalies in the case of nasal vowels. $AbnRecall$ measures computed for this category is 0.77 (strategy 2) compared to 0.83, 0.84 and 0.86 values associated with for oral vowels, voiced and unvoiced consonants respectively. This may be due to the limited data used in the model training of this category. Surprisingly, this category presents the best $AbnPrec$ ratio with a 0.77 value, close to oral vowels (0.75), but far from 0.57 and 0.61 values associated with voiced and unvoiced consonants respectively.

On the basis of these first results, some assumptions/observations can be highlighted :

— the behavior of the automatic systems, and notably the SVM-based one, is quite stable, comparing performance measures between female and male speech productions (confirmed along the different speech recordings available per speaker), discarding the ”male 1” from assumption ;

— both systems show better results on severe dysarthric patients. For speakers ”female 2”, ”female 4” and ”male 2” who are considered as the most dysarthric patients, SVM-classification system reaches almost the best $AbnRecall$ and $AbnPrec$ measures, (the best values : 1 and 0.98 respectively for patient ”female 2”) while patients ”male 1”, ”female 1” and ”female 3” who are considered as the least dysarthric patients show lower performance for both measures ;
— the set of features used in the SVM-classification system as well as its discrimination process shows relevancy for the abnormal phoneme detection, increasing the recall values compared with the baseline system. Nevertheless, even though some improvements are observed regarding the precision measurements, these values still remain rather low, demonstrating that the automatic system tends to falsely detect too many abnormal phonemes and, therefore, to be more severe than the human expert ;
— these observations are supported by the agreement rates computed on the control speakers. Both systems reject an average of 9% of phonemes, considered as abnormal on female speakers whereas the SVM-classification system outperforms the baseline system on men speakers, by rejecting 4% phonemes only compared to the 7% rejected by the baseline approach. Even though, one suppose that the anomalies detected by both automatic systems on control speakers are false positive errors, it cannot be excluded that some or all of them are, in fact, some true positive errors since abnormalities could exist within speech produced by a healthy speaker ;
— differences in the SVM-based system behavior can be noted comparing performance reached on the individual phonetic categories. Indeed, if the recall measures can be considered as similar between categories, precision measures show that the system tends to falsely detect more anomalies within consonants than vowels.

4.2. Corpus 2

The first corpus used in this study contains only 8 patients (even though each one made more than one recording), all suffering from the same disease (lysosomal storage disease). Moreover, the human annotation of the abnormal phonemes was carried out by a single expert, which makes it unreproducible due to its subjective nature. Consequently, the experimental evaluation of the automatic anomaly detection based on automatic approaches is extended here to the second database, corpus 2, described in section 3.1.2. Nevertheless, as no specific annotation regarding the normality of phoneme production was made on this corpus, the relationship between the rate of phonemes detected as abnormal by the SVM-based classification system (over the amount of available phonemes) and the results of the perceptual evaluation performed by the expert jury is studied here.

Figures 2 and 3 depict the rate of abnormal phonemes detected automatically on each patient relatively to his/her dysarthria severity degree grouped by women and man respectively. The couple of figures show an important relationship between both measures, confirmed by an overall correlation ratio of 0.89 and 0.86 for women and men respectively. This observation, even though not conclusive, tends to support the behavior of the automatic system previously observed on corpus 1, for which the recall and precision measurements tend to increase for higher dysarthria severity degrees. Therefore, the high correlation observed here, even if it does not prove the accuracy of the automatic anomaly detection approach at the phoneme level, seems to confirm the potentiality of the system to deal with acoustic speech alterations.

The perceptual evaluation performed by the expert jury covering other speech quality criteria, similar comparison results are presented, per disease, in table VII. These figures reveal interesting differences between diseases and genders along with the different criteria :

Fig. 2. Relationship between the SVM-classification based abnormality rate and the dysarthria severity degree for female speakers issue from corpus 2

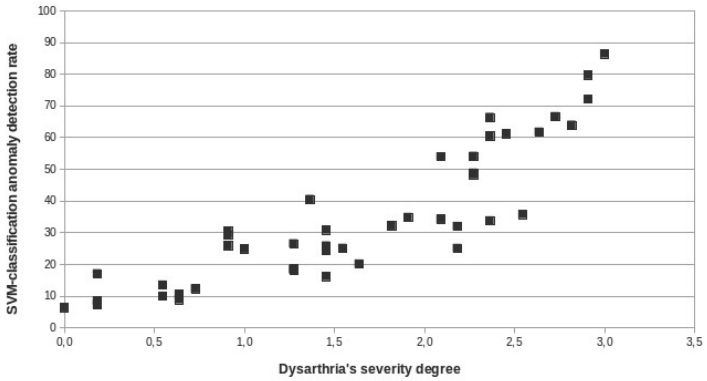
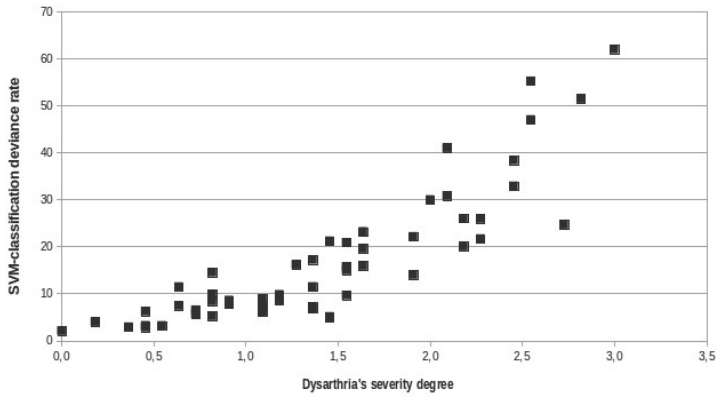


Fig. 3. Relationship between the SVM-classification based abnormality rate and dysarthria severity degree for male speakers issued from corpus 2



— high correlations are achieved with the global severity degree of dysarthria for all diseases (near to or above 0.9) except for patients suffering from cerebellar ataxia, and notably the female speakers,
— similar behavior is observed with both the articulation impairment and intelligibility degree items, even values are slightly lower (between 0.8 and 0.9),
— regarding the speech rate, values are more heterogeneous and contrasting. There is no real trend depending on gender or disease. These observations tend to be consistent with the design of the automatic anomaly detection approach. Indeed, if the design of the automatic system is expected to be strongly correlated with the loss of intelligibility or the articulation impairment, the definition of the feature set used to characterize phonemes and the level of detection (phoneme) itself are hardly compatible with the consideration of the speech rate. Only information on the phoneme length, present in the feature set, could reflect this criterion.

Table VII. Correlation between the rate of abnormal phonemes detected automatically and perceptual evaluation measures (averaged over the different experts) associated with corpus 2.

Disease	Gender	Global severity degree	Articulation impairment	Intelligibility	Speech rate
Parkinson's disease	Female	0.89	0.86	0.89	0.81
ALS	Female	0.91	0.86	0.83	0.92
Cerebellar ataxia	Female	0.52	0.5	0.38	0.64
Lysosomal storage disease	Female	0.9	0.81	0.87	0.43
Parkinson's disease	Male	0.87	0.8	0.84	0.6
ALS	Male	0.91	0.82	0.86	0.67
Cerebellar ataxia	Male	0.81	0.65	0.83	0.59
Lysosomal storage disease	Male	0.96	0.88	0.68	0.69

Table VIII. Performance of SVM-based Classification approaches applied on the control speakers issued from corpus 2, expressed in terms of agreement rate measure (%)

Patients	Global dysarthria severity degree	SVM-based classification system
Control men average	0.16	97.1
Control women average	0.10	89.6
Control speaker average	0.13	93.35

— Finally, it is interesting to point out that patients with Parkinson's disease, ALS or lysosomal storage disease display similar correlation rates. This may sustain that the very encouraging performance of the SVM-based classification system reached on corpus 1 (sharing the same speech productions of patients suffering from lysosomal storage disease as corpus 2) could be transposable to other pathologies.

Finally, table VIII presents the SVM-based classification system performance when applied on the control speakers of corpus 2, expressed in terms of agreement rate. The system considers as abnormal about 10% and 3% phonemes for women and men speakers respectively. These results match nearly perfectly those observed on corpus 1 (an average of 9% and 4% phonemes rejected for women and men respectively, see table V).
To summarize, these last observations, in addition to the high correlation ratio observed with the perceptual assessment tend to support the approach consistency whenever applied on dysarthric or normal speech productions.

5. CONCLUSION

This paper investigates a novel approach for detecting acoustic anomalies in speech signal produced by patients suffering from dysarthria. The main advantage of this novel approach, compared to previous work done by the authors on the basis of a simpler detection system, is to model both normal and abnormal parts of speech manually annotated by an expert. The performance evaluation of the proposed approach is carried out on two different corpora, both comprising speech productions recorded from patients suffering from dysarthria and healthy speakers. The first one, based on labeled data for both normal and abnormal part of speech, highlights very promising results, notably in terms of matching between the automatic detection of acoustic alterations and annotations of the human expert. Nevertheless, this evaluation demonstrates also that the automatic system tends to be more severe than the human expert by detecting too many anomalies. Even, very interesting results have been observed by focusing on more specific phonetic production, like vowels, voiced and unvoiced consonants or by comparing results between male or female speakers. Though, this first corpus presents some limits: it was perceptually evaluated by a unique human expert in order to provide labeled data including anomalies in speech signal and such a

perceptual evaluation, even when performed by an experienced professional, remains subjective and subject to critics. Secondly, even though different speech recordings are available per speaker, the corpus comprises 8 patients only, suffering from a rare disease linked to a mixte dysarthria. For these reasons, the automatic approach was evaluated on a second corpus of speech produced by patients with Parkinson's disease, ALS or cerebellar ataxia. Instead of labeled data as normal or abnormal, this corpus was annotated by a jury of 11 experts on several perceptual criteria. Since no labeled data regarding anomalies is available, evaluation was carried out by observing the relationship between the rate of speech parts annotated as abnormal by the automatic approach and the perceptual rates given by the expert jury regarding the global severity degree of dysarthria, the intelligibility measure, the articulation impairment and the speech rate of speakers. Result analysis points out very interesting behavior of the automatic system exhibiting some very relevant correlation rates with the major of perceptual criteria and supporting the assumption that performance reached by the automatic approach in the task of the anomaly detection on the first corpus could be transposable to other kind of diseases like Parkinson's disease or SLA present in the second corpus.

In future work, attention will be made on different directions. First of all, investigation will be made in order to analyze further the behavior of the automatic approach, especially in regards to its lower values of precision rates. Indeed, it will be interesting to examine why the automatic system tends to be more severe than the human expert, and notably on consonants as observed in this paper. Besides, it would be also relevant to determine whether the automatic system is more relevant on certain acoustic alterations (voicing, spectral distorsion, ...) than on others. The latter may open a larger research question : How is the relationship between the human perception of alterations in speech and their modeling by automatic speech processing systems ?

ACKNOWLEDGMENTS

This work has been carried out thanks to the support of the BLRI Labex (ANR-11-LABEX-0036) and the A*MIDEX project (ANR-11-IDEX-0001-02) funded by the "Investissements d'Avenir" French government program managed by the ANR and thanks to the French ANR projet Typaloc (ANR-12-BSH2-0003-03). We deeply thank Georges Linares for his help regarding the use of the phonetic alignment tool.

REFERENCES

F. Brugnara, D. Falavigna, and M. Omologo. 1993. Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models. *Speech Communication* 12(4) (1993), 357–370.

V. Chandola, A. Banerjee, and V. Kumar. 2007. *Anomaly detection : a survey*. University of Minnesota (US).

H. Christensen, P. Green, and T. Hain. 2013. Learning speaker-specific pronunciations of disordered speech. In *Proceedings of Interspeech'13*. Lyon, France.

S. Lilly Christina, P. Vijayalakshmi, and T. Nagarajan. 2012. HMM-based speech recognition system for the dysarthric speech evaluation of articulatory subsystem. In *International Conference on Recent Trends In Information Technology (ICRTIT)*.

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

J. R. Duffy. 2005. *Motor speech disorders: substrates, differential diagnosis and management*. Motsby- Yearbook, St Louis, 2e dition.

P. Enderby. 1983. Frenchay dysarthric assessment. *Pro-Ed, Texas* (1983).

C. Fougeron, L. Crevier-Buchman, C. Fredouille, A. Ghio, C. Meunier, C. Chevrie-Muller, J.-F. Bonastre, A. Colazo-Simon, C. Delooze, D. Duez, C. Gendrot, T. Legou, N. Lvque, C. Pillot-Loiseau, S. Pinto, G. Pouchoulin, D. Robert, J. Vaissire, F. Viallet, and C. Vincent. 2010. The DesPho-APaDy Project: Developing an Acoustic-phonetic Characterization of Dysarthric Speech in French. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (19-21). Valletta, Malta.

C. Fredouille and G. Pouchoulin. 2011. Automatic detection of abnormal zones in pathological speech. In *Intl Congress of Phonetic Sciences (ICPHs'11)*. Hong Kong.

S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier. 2005. ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. In *Proceedings of Interspeech'05*. 1149–1152.

A. Ghio, G. Pouchoulin, B. Teston, S. Pinto, C. Fredouille, C. De Looze, D. Robert, F. Viallet, and A. Giovanni. 2012. How to manage sound physiological and clinical data of 2500 dysphonic and dysarthric speakers? *Speech Communication* 54(5) (2012), 664–679.

K. C. Hustad. 2008. The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of Speech, Language and Hearing Research* 51(3) (2008), 562–573.

T. Joachims. 1999. Making large-Scale SVM Learning Practical. In *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola (Eds.). MIT Press, Cambridge, MA, Chapter 11, 169–184.

R. D. Kent, G. Weismer, J. F. Kent, H. K. Vorperian, and J. R. Duffy. 1999. Acoustic studies of dysarthric speech: Methods, progress, and potential. *The Journal of Communication Disorders* 32:3 (1999), 141–186.

L. Lhoussaine. 2012. *Premire validation de la Grille dEvaluation Perceptive de la Dysarthrie (G.E.P.D.) : effet du niveau dexpertise du jury et diffrenciation entre types de dysarthrie*. Ph.D. Dissertation. Speech therapist thesis, University of Paris VI, Pierre er Marie Curie (in French).

A. Lowit and R. D Kent. 2010. *Assessment of motor speech disorders*. Vol. 1. Plural publishing.

J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel. 1999. Performance measures for information extraction. *Proceedings of DARPA Broadcast News Workshop* (1999).

C. Middag, J.-P. Martens, G. Van Nuffelen, and M. De Bodt. 2009. Automated Intelligibility Assessment of Pathological Speech Using Phonological Features. *EURASIP Journal on Applied Signal Processing* (2009).

B. E. Murdoch (Ed.). 1998. *Dysarthrie : a physiological approach to assessment and treatment*.

K. MyungJong and K. Hoirin. 2012. Automatic Assessment of Dysarthric Speech Intelligibility Based on Selected Phonetic Quality Features. In *Computers Helping People with Special Needs*. Lecture Notes in Computer Science, Vol. 7383. 447–450.

M. Parker, S. Cunningham, P. Enderby, M. Hawley, and P. Green. 2006. Automatic speech recognition and training for severely dysarthric users of assistive technology: the STARDUST project. *Clinical Linguistics and Phonetics* 20(2–3) (2006), 149–156.

K. M. Rosen and J. R. Duffy R. D. Kent, A. L. Delaney. 2006. Parametric quantitative acoustic analysis of conversation produced by speakers with dysarthria and healthy speakers. *Journal of Speech, Language, Hearing Research* 49(2) (2006), 395–411.

K. M. Rosen and S. Yampolsky. 2000. Automatic speech recognition and a review of its functioning with dysarthric speech. *Augmentative and Alternative Communication* 16(1) (2000), 48–60.

F. Rudzicz, A. K. Namasivayam, and T. Wolff. 2012. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC'12)*. 523–541.

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

H. Strik, E. Sanders, M. Ruiter, and L. Beijer. 2002. Automatic recognition of dutch dysarthric speech: a pilot study. In *Proceedings of International Conference on Spoken Language Processing (ICSLP'02)*. Denver, US, 661–664.

B. Tomik and J.R. Guiloff. 2010. Dysarthria in amyotrophic lateral sclerosis: a review. *Amyotrophic Lateral Sclerosis* 11 (1–2) (2010), 4–15.

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

K. M. Yorkston, E. Strand, R. Miller, A. Hillel, and K. Smith. 1993. Speech deterioration in amyotrophic lateral sclerosis: im- plications for the timing of intervention. *Journal Med Speech Language Pathology* 46 (1993), 35–46.