AUTOMATIC DETECTION OF ABNORMAL ZONES IN PATHOLOGICAL SPEECH

Corinne Fredouille, Gilles Pouchoulin

University of Avignon, CERI/LIA
(corinne.fredouille, gilles.pouchoulin)@univ-avignon.fr

ABSTRACT

This paper proposes an original methodology devoted to the automatic detection of abnormal zones in speech utterances in the specific context of impairments. This methodology relies on automatic speech processing, involving an automatic text-constrained phoneme alignment, the computation of phoneme based-normalized acoustic scores and a reference scale, permitting to label *in fine* a phoneme as normal or abnormal. The evaluation of the methodology reliability when applied to a dysarthric speech corpus has shown very encouraging results, highlighting the efficiency of the methodology in detecting true abnormal zones. In addition, this evaluation underlines a lack of precision (in terms of "information retrieval") resulting in mis-labelled normal zones in the automatic annotation (false positive), leading to further investigation.

Keywords: Speech disorders, dysarthria, automatic speech processing, detection of abnormal zones

1. INTRODUCTION

Dysarthria is a motor speech disorder due to damages of the nervous system. Many studies have been proposed in the literature to characterize dysarthria acoustically and/or to propose a classification of dysarthria [1, 2, 3]. Although the main features that differentiate "typical" patients affected by different types of dysarthria have been identified, the study of dysarthria needs more comprehensive phonetic descriptions to encompass the great diversity observed in patients' speech patterns.

Automatic speech processing-based tools are largely used in the literature to deal with dysarthric speech. Mainly, the goal of such approaches is to provide patients with assistive technologies [4] or to provide technologies for an objective assessment of the dysarthric speech severity [5]. In this paper, the authors propose an original approach to detect automatically abnormal zones of speech in the acoustic signal. This approach based on speech processing tools aims to help phoneticians in their manual analysis of dysarthric speech production by guiding them towards acoustic zones potentially attractive for subsequent fine grained acoustic investigation. It also permits to deal with larger patient corpora as well as longer speech production as only a limited set of speech zones will be targeted in the signal.

2. ABNORMAL ZONE DETECTION

The methodology proposed in this paper to detect automatically abnormal zones in speech utterance relies on three main steps: (1) an automatic phoneme alignment, (2) the computation of normality scores at the phoneme level, and (3) the design of abnormality zone cartography. The following subsections are dedicated to each of these steps.

2.1. Automatic phoneme alignment

The first step of the detection methodology is to segment speech utterances into fixed zones, which will be analyzed further to determine whether they have to be considered as normal or abnormal. Here, the phoneme level has been chosen because (1) their duration is considered as sufficient to provide usable normality scores (see next subsection), notably compared with the frame level, (2) they could be acoustically distorted due to articulatory impairments.

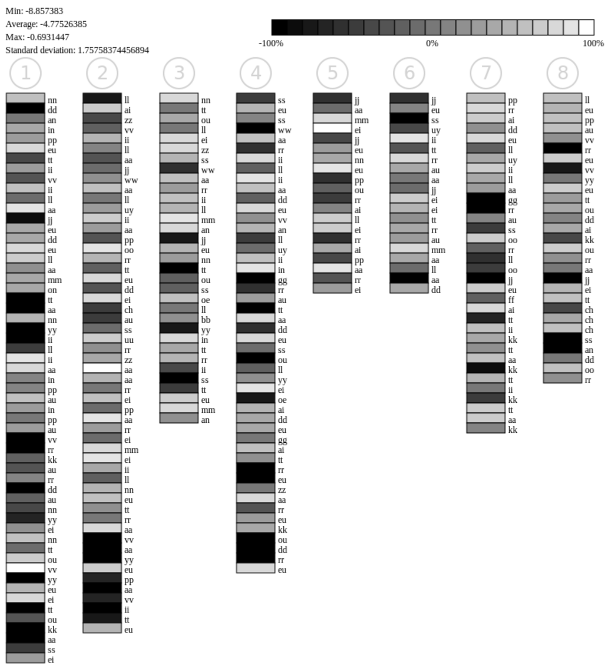
The segmentation of speech utterances is carried out by an automatic text-constrained phoneme alignment tool. In this sense, it takes as input the sequence of words pronounced in each speech utterance, via an orthographic transcription performed by human listeners. Additionally, it has as input a restricted lexicon of words associated with some phonological variants, based on a set of 38 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 [6] for more details).

The speech segmentation results in a couple of start and end boundaries per phoneme present in the orthographic transcription.

2.2. Acoustic score measurement

Once the automatic text-constrained alignment is applied as reported in the previous subsection, each speech utterance can be coupled with the set of

Figure 1: Example of cartography relating to a female patient, illustrating the degree of normality per phoneme



phonemes pronounced and their temporal positions in the speech signal (start and end boundaries). In this context, the goal of this step is to associate a normalized acoustic score with each of these phonemes, which will permit further to determine a normality degree. In this paper, this normalized acoustic score is defined as follows: $L_p^{norm}(y_p) = log(\frac{L_p^{Constrained}(y_p)}{L_{p'}^{Unconstrained}(y_p)})$ where $L_p^{norm}(y_p)$ is the expected normalized acoustic score computed for a given phoneme p on the related speech signal y_p. $L_p^{Constrained}(y_p)$ is an acoustic score assigned to phoneme p by the automatic text-constrained phoneme alignment process. $L_{p'}^{Unconstrained}(y_p)$ is an acoustic score assigned to phoneme p', potentially different from phoneme p. This score is set by comparing the acoustic scores obtained by all the phonemes available (included p) and by choosing the maximum one (associated with phoneme p'). In this context, either p and p' denote the same phoneme and the normalized acoustic score will be equal to 0, either p and p' are different and the normalized acoustic score will tend towards negative values according to the acoustic distortion of phoneme p.

2.3. Cartography

The last step of the methodology consists in exploiting the normalized acoustic scores assigned individually to each phoneme. This exploitation consists in determining if a phoneme has to be considered as normal or abnormal from the speech production point of view, especially in the context of speech disorders. The degree of normality of a given phoneme is determined according to a reference scale. The latter takes into account all the normalized acoustic scores computed from the set of control speakers, fixing the minimum, maximum, and median values over all the phonemes. Thus, analyzing the normalized acoustic scores related to the speech utterance of a given patient will permit to observe whether phonemes are inside or outside the reference scale.
In order to facilitate this observation, the normalized acoustic scores can be easily represented graphically, on a "normality cartography". One example of cartography is provided in figure 1. On this illustration, the reference scale is reported on the top-right, representing the degree of normality by a grey color gradation: the blacker the color is, the more abnormal the phoneme is. Each column represents the sequence of phonemes - themselves represented by 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 which zones of the speech signal are considered as abnormal by the automatic processing and to know which phonemes are associated with. This underlines the relevancy of the methodology proposed. The next sections will be dedicated to the evaluation of the reliability of both normalized acoustic scores and reference scale associated with.

3. METHODOLOGY EVALUATION

In order to be able to evaluate and measure the reliability of the automatic speech processing-based methodology described in the previous section, it is necessary to determine the real abnormal zones in a speech utterance. As it does not exist other automatic process able to perform this task, we have to confront the results of this automatic methodology with an human expertise. The latter has to label perceptually speech zones as normal or abnormal by listening to and analyzing the speech signal analysis. This permits to compute different agreement measures between (normal and abnormal) labels coming from the automatic methodology and the human expert.
We propose different measures, some of them stem from the retrieval information domain [7]:
• the correct agreement rate regarding the detection of both normal and abnormal zones (in %), named in the rest of the paper $AggRate$, given by the ratio between the number of zones well labeled by the

automatic processing (according to the human expertise) and the total number of zones;

• the abnormality class-based recall measure (value between 0 to 1), named $AbnRecall$, given by the ratio between the number of zones well detected as abnormal 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 correct abnormal zones. The more close to 1 the ratio is, the more the automatic system performs well to detect real abnormal zones;

• the abnormality class-based precision measure (value between 0 to 1), named $AbnPrec$, given by the ratio between the number of abnormal zones well detected by the automatic processing and the number of zones that the automatic processing labels as abnormal (truly or falsely). This ratio will measure the inverse rate of false alarm produced by the automatic methodology in detecting the abnormality zone: the more close to 1 the ratio is, the more precise in detecting the abnormal zones the automatic system is.

It is worth noting that:

(1) both $AbnRecall$ and $AbnPrec$ are complementary in order to evaluate the reliability of the automatic processing in detecting abnormal zones;

(2) the reliability of the methodology, while applied on the control speakers, can be only evaluated through the $AggRate$ measurement (Null value for $AbnRecall$ and $AbnPrec$).

4. EXPERIMENTS

The automatic speech processing-based methodology proposed for the detection of abnormal speech zones is applied and evaluated on dysarthric speech. This evaluation will involve the measurements described in the previous section.

4.1. Corpus

The study carried out in this paper is based on a dysarthric speech corpus, recorded at the hospital La Pitié-Salpétrière in Paris. The corpus is composed of recordings of 7 control speakers and 8 patients. Patients suffer from rare lysosomal storage diseases and show disparities in the severity degree of dysarthria according to the progression of their disease. All the speakers were recorded longitudinally: all six months approximatively for 2 years for the patients (resulting in 3 to 5 recording sessions each), each week for 1 month for the control speakers (3 to 5 recording sessions each).

All the speakers were recorded in similar conditions, reading a French fairytale called “Le cordonnier” (The cobbler). The duration of speech utterances varies from 48s to 196s, with an average of around 60s for control speakers, and 85s for patients.

All the speech utterances related to the patients were analyzed by an human expert in order to annotate the abnormal speech zones. Helped with the listening and the Praat-based analysis of the speech signal coupled with the automatic phoneme segmentation, the precise task of the expert was to label a phoneme as normal or abnormal, by indicating in this last case, the type of abnormality (noise, voicing impairment, spectral distortion, ...).

4.2. Results

The methodology described in section 2. is applied on the dysarthric patients and control speakers. This results in a set of phoneme-based normalized acoustic scores per speaker as well as their corresponding values on the reference scale. The comparison with the annotation of the human expert permits to compute measures proposed in section 3., reported in tables 1 and 2 for patients and control speakers respectively (averaged measures over the different recording sessions related to each speaker). This comparison is carried out following two approaches :

• App1: considering the phoneme uniquely to evaluate the decision of the automatic methodology (first value reported in table 1);

• App2: considering the phoneme as well as the previous and next phonemes to evaluate the decision of the automatic methodology in detecting abnormal zones (second value reported in table 1). In this case, if the human expert considers a given phoneme as abnormal while the automatic methodology detects the previous or last phoneme as abnormal (but not the given phoneme), then a good match is notified. This approach aims to support a one phoneme-based shift in the automatic detection due, for instance, to short alignment shifts. This approach has no incidence on the measures relating to the control speakers.

Finally, table 1 reports per speaker the averaged percentage of abnormal phonemes annotated by the human expert, considering all the recording sessions.

Observing $AggRate$ measures, the automatic methodology obtains averages of 68% for the female patients, and 75% for the male patients considering the phoneme only (App1), and averages of around 85% for both the female and male patients considering the phoneme plus its context (App2). In addition, an average of around 92% is reached for both the female and male control speakers. If these results are quite encouraging, it is interesting to notice that the values per speaker are more homogeneous

Table 1: Performance of the automatic methodology applied on the dysarthric patients, expressed in terms of correct agreement rate ($AggRate$), abnormality class-based recall measure ($AbnRecall$) and precision measure ($AbnPrec$). Value pairs X/Y denote the computation of measures at the phoneme level only (X) and at the phoneme level considering its context (Y)

Patients (% abnormal zones)	$AggRate$ (in %)	$AbnRecall$ ($[0, 1]$)	$AbnPrec$ ($[0, 1]$)
Male 1 (6,7)	88,0 / 89,6	0,10 / 0,32	0,10 / 0,25
Male 2 (33,9)	66,2 / 82,4	0,40 / 0,77	0,40 / 0,70
Male 3 (15,6)	79,5 / 85,5	0,47 / 0,75	0,27 / 0,50
Male 4 (26,6)	68,0 / 79,3	0,44 / 0,67	0,35 / 0,58
Female 1 (15)	82,0 / 87,6	0,5 / 0,77	0,31 / 0,52
Female 2 (23,5)	70,0 / 82,0	0,6 / 0,89	0,35 / 0,55
Female 3 (13,1)	75,5 / 82,0	0,49 / 0,76	0,21 / 0,42
Female 4 (78,5)	47,0 / 87,7	0,43 / 0,87	0,66 / 0,96

Table 2: Performance of the automatic methodology applied on the control speakers, expressed in terms of correct agreement rate regarding the detection of both normal and abnormal zones ($AggRate$)

Control	$AggRate$
Male 1	93,3
Male 2	94,2
Male 3	87,9
Female 1	92,4
Female 2	90,8
Female 3	92,3
Female 4	93,4

considering the App2 measurement approach.
Observing $AbnRecall$ and $AbnPrec$ measures, the automatic methodology obtains averages of 0,51 and 0,38 for the female patients respectively, and 0,36 and 0,27 for the male patients respectively considering the phoneme only (App1), and averages of 0,82 and 0,61 for the female patients respectively, and 0,63 and 0,52 for the male patients respectively considering the phoneme plus its context (App2). From these results, it can be pointed out that the automatic methodology reaches promising measures considering the App2 approach, even on the male patients for which only one patient (male 1) contributes to the decrease of values. The high $AbnRecall$ measures obtained by most of the patients indicate that the automatic methodology is able to detect efficiently the abnormal zones highlighted by the human expert. In contrast, moderate $AbnPrec$ values obtained by most of the patients indicate that the automatic methodology tends to over-detect abnormal zones, leading to some noise (false positive) in the labeling. This is supported by the 8% detection errors reported on the control speakers. Efforts in terms of understanding of the types of errors made by the automatic methodology on the control speakers have to be done in order to minimize them. This understanding will permit to increase the $AbnPrec$ measures on the patient speech utterances.

5. CONCLUSION

In this paper, an original methodology based on the automatic speech processing is proposed in order to detect abnormal zones in pathological speech. An evaluation protocol is also proposed, which has shown very encouraging results when the methodology is applied on a dysarthric speech corpus.
Further work will be dedicated in improving the automatic methodology, especially regarding the precision measures, which remain relatively low compared with the recall measures. A second investigation will be carried out in order to apply the methodology on other speech dysarthric corpora or another impairment context like the dysphonia.

6. ACKNOWLEDGMENT

This work has been funded by the French National Research Agency (ANR) in the context of the DesPho-APady project (ANR BLAN08-0125). We deeply thank Nathalie Lévèque et Frédéric Sedel to provide us with the dysarthric corpus, Olavo Panseri - our human expert - for his manual annotation of the corpus as well as Cécile Fougeron for her help.

7. REFERENCES

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

[2] F. L. Darley, A. E. Aronson, and J. R. Brown, *Motor speech disorders*, Philadelphia, 1975.

[3] R. D. Kent, J. F. Kent, J. R. Duffy, and G. Weismer, "The dysarthrias: Speech-voice profiles, related dysfunctions, and neuropathologies," *Journal of Medical Speech-Language Pathology*, pp. 165–211, 1998.

[4] S. K. Fager, D. R. Beukelman, T. Jakobs, and J.-P. Hosom, "Evaluation of a speech recognition prototype for speakers with moderate and severe dysarthria: A preliminary report," *Augmentative and Alternative Communication*, vol. 26(4), pp. 267–277, December 2010.

[5] C. Middag, J.-P. Martens, G. Van Nuffelen, and M. De Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Applied Signal Processing*, 2009.

[6] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on hidden markov models," *Speech Communication*, vol. 12(4), pp. 357–370, 1993.

[7] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," *Proceedings of DARPA Broadcast News Workshop*, 1999.