AUTOMATIC SPEECH PROCESSING FOR DYSARTHRIA: A STUDY OF INTER-PATHOLOGY VARIABILITY

Imed Laaridh[⋆†], Corinne Fredouille[⋆], Christine Meunier[†]

[⋆]University of Avignon, CERI/LIA, France [†]University of Aix Marseille, CNRS, LPL UMR 7309, 13100, Aix-en-Provence, France

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

ABSTRACT

Despite their large advances, especially for consumer applications, automatic speech technologies still encounter very huge difficulties when they are exposed to dysarthric speech. However, they were presented very early as potential solutions to provide objective tools to deal with speech disorders in order to help clinicians in their clinical practice and patients in their everyday life. In order to understand the difficulties encountered by automatic speech processing, this paper investigates the reliability of a simple automatic phone alignment when dealing with dysarthric speech. Notably, the corpus used involves French read speech recordings produced by patients suffering from four different pathologies, exhibiting three different types of dysarthria. The observations of the segmentation outputs yielded by the automatic tool (compared with a manual segmentation) according to the pathologies, the type of dysarthria and different phonetic categories reveal a very large heterogeneity of behavior between pathologies, and within a same pathology.

Keywords: Dysarthria, automatic text-constrained phone alignment, pathology-dependent phonetic analysis

1. INTRODUCTION

Dysarthria is a motor speech disorder, consequence of 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 the speech production, like respiratory, phonatory, resonatory, articulatory and prosodic. Consequently, this may be reflected by weakness, spasticity, incoordination, involuntary movements, or abnormal muscle tone, depending on the location of the neurological damage. Research on dysarthria is very abundant and has covered different axes for about fifty years : perceptual evaluation of speech alterations for dysarthria classification [5, 6, 7], perceptual measurement of dysarthria severity, notably related to the speaker's intelligibility ([8, 26, 13, 16]), articulatory or/and acoustic analysis ([14, 17, 21, 4, 11, 25]), automatic speech processing for intelligibility assessment ([2, 18, 19, 15]) or for speech recognition ([22, 12, 20, 23, 3]). The aim of these studies has been to better understand and characterize diseases and related speech disorders, to help clinicians for diagnosing and following the condition progression of patients, but also to design therapy objectives, or to help patients with severe speech disorders in their everyday life through Alternative and Augmented Communication (AAC) tools.

Regarding the automatic approaches, they were presented very early as potential solutions to provide objective tools to deal with speech disorders [9]. However, if speech technologies have reached an advanced stage for consumer applications, major issues remain for dysarthric speech application. Even though there is a set of typical acoustic-perceptual cues, such as imprecise consonants, vowel centralization, slow rate, monopitch, monoloudness, hypernasality, that is commonly used to characterize the main disturbances of the various types of dysarthria in the 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 and also within the same group [24]. The objective of the work[1] presented in this paper is to observe and analyze the behavior of a very straightforward speech processing task when it is applied on a dysarthric speech corpus. The particularity of this study is to involve a speech corpus produced by French patients suffering from four different diseases and types of dysarthria. The speech processing task relies on an automatic text-constrained phone alignment, which "simplicity"[2] will face potential difficulties in dealing with dysarthric speech.

2. AUTOMATIC SEGMENTATION SYSTEM

An automatic text-constrained phone alignment tool aims at providing a segmentation of speech utterances into phones. This tool takes as input the sequence of words pronounced in each utterance and a phonetized lexicon of words. The latter is based on a set of 37 French phones and includes different phonological variants per word. The se-

quence of words comes from an orthographic transcription performed by a human listener. This manual transcription follows some specific rules to denote deletions, substitutions, insertions and repetitions of some words and/or phone sequences produced by speakers. Moreover, for this manual transcription, speech records are split into inter-pausal units (IPUs). An IPU is defined as a pause-free unit of speech from the same speaker separated from another IPU by at least 250ms of silence.

The automatic alignment process is based on a Viterbi decoding and graph-search algorithms, the core of which is the acoustic modeling of each phone based on a Hidden Markov Model (HMM). The HMM-based models used in this work are built using the Maximum Likelihood Estimate paradigm on the basis of about 200 hours of French radiophonic speech recordings [10]. In order to get speaker-dependent models, a three-iteration Maximum A Posteriori (MAP) adaptation is performed to all the HMM parameters. Finally, acoustic vectors consist of 12 Perceptual Linear Prediction coefficients plus the energy, plus their delta and delta-delta coefficients. This automatic alignment process results in a couple of start and end boundaries per phone produced in the speech records.

3. EXPERIMENTAL PROCEDURE

3.1. Corpus

The current study is based on a speech corpus produced by 25 speakers: 13 healthy speakers (control) and 12 dysarthric patients. The patients suffer from various diseases: Amyotrophic Lateral Sclerosis (ALS), Parkinson's Disease (PD), Cerebellar Ataxia (CA) and Lysosomal Storage Disease (LSD) and present various dysarthria severity degrees (DSD).

All the participants were asked to read the same text, a French fairy-tale called "Le cordonnier" (The cobbler), as naturally as possible. In this paper, we however use only the first paragraph of this text containing 215 phones (95 vowels and 120 consonants). The duration of speech utterances varies from 21s to 61s with an average of about 26s for the control speakers and 32s for the patients. The differences in duration observed for the patients are due to differences in speech rate.

All the speech recordings were evaluated perceptually by a jury of 11 experts. They were asked to rate all the patients on perceptual items of speech quality. These items included the global evaluation of the Dysarthria Severity Degree (DSD) rated on a scale from 0 to 3 (0 -no dysarthria, 1 -mild, 2 -moderate, 3 -severe dysarthria) and the evaluation of speech rate on a scale from -3 to 3 (-3 -very slow, 0 -normal, 3 -extremely fast speech rate) on which this paper is focused.

In addition, all the speech utterances were analyzed by human experts based on their listening and the Praat [1] tool in order to provide a manual segmentation of phones. This manual segmentation was carried out by making corrections of phone boundaries, if necessary, on the basis of an automatic phone segmentation. It has to be noted that the human expert could encounter difficulties in defining phone boundaries, especially when the speech quality was dramatically altered. Such non-segmentable phone sequences were not considered in the rest of the paper. Table 1 provides detailed information on this speech corpus, including per disease the number of patients, the minimum and maximum DSD and the minimum and maximum speech rate. We can point out that the speech corpus used in this paper is produced by patients with mild dysarthria.

3.2. Evaluation

In order to analyze the behavior of the text-constrained phone alignment tool, the automatic phone segmentation outputs are compared to the manual outputs provided by the human expert. In this context, three measures, expressed in ms, are computed for each phone :
- the Start Shift (*SS*), which is given by the absolute value of the difference between the phone start boundaries from the automatic and manual segmentations;
- the Midpoint Shift *MS*, which is given by the absolute value of the difference between the phone midpoints from the automatic and the manual segmentations;
- the Duration Difference *DD*, which is given by the difference between the phone durations from the automatic and manual segmentations (negative values correspond to shorter durations of phones from the automatic segmentation).

4. RESULTS AND DISCUSSION

Table 1: Information related to the corpus: the number of speakers, the minimum and maximum Dysarthria Severity Degrees (DSD) and the minimum and maximum speech rate per disease.

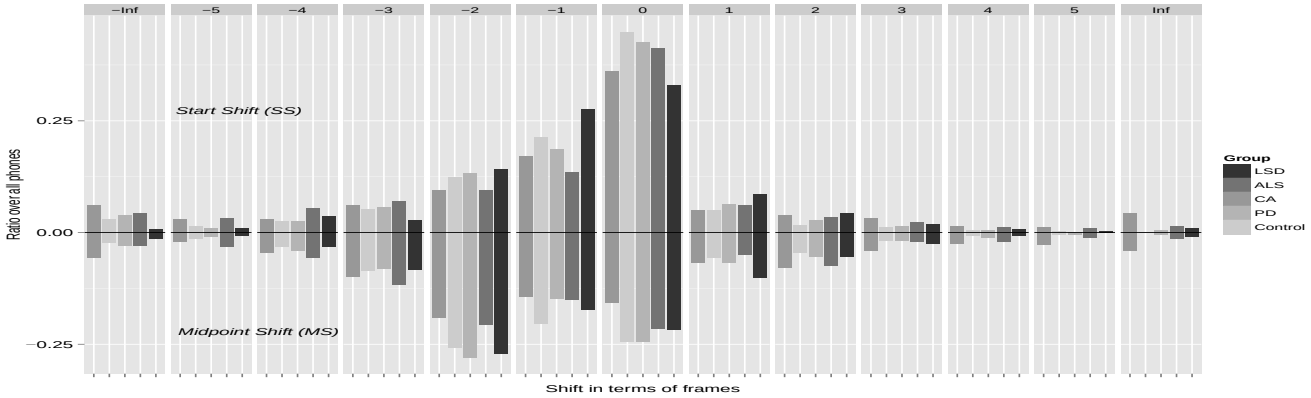
Disease	Number of speakers	(Min,Max) DSD	(Min,Max) speech rate
ALS	2	(0.9,1.5)	(-0.5,-0.2)
PD	4	(0.4,1.3)	(-0.1,1.7)
CA	4	(0.9,1.5)	(-2.2,-0.8)
LSD	2	(1.5,1.5)	(1.5,2.8)
Control	13	-	-

This section details and discusses the behavior of the automatic phone segmentation tool on both nor-

Table 2: Performance of the automatic phone segmentation system expressed in terms of average and standard deviation (σ) of Start Shift *SS*, Midpoint Shift *MS* and Duration Difference *DD* given per pathology and phonetic category. All the measures are expressed in terms of ms.

Phonetic category	Measure	Control speakers	Pathology			
			LSD	ALS	CA	PD
Unvoiced Plosives	*SS* (σ)	11.2 (21.2)	17.6 (32.7)	13.0 (13.1)	16.1 (27.5)	11.3 (24.9)
	MS (σ)	10.7 (13.0)	11.9 (17.6)	9.2 (8.9)	14.2 (16.7)	9.8 (14.2)
	DD (σ)	-3.5 (29.2)	-7.7 (21.4)	-5.6 (22.2)	-5.4 (32.0)	-1.8 (15.5)
Voiced Plosives	*SS* (σ)	8.5 (14.2)	13.9 (14.2)	24.6 (59.3)	21.9 (28.6)	13.1 (23.4)
	MS (σ)	6.9 (9.5)	12.8 (19.2)	21.3 (33.2)	18.0 (21.7)	10.5 (13.0)
	DD (σ)	4.7 (22.4)	15.4 (38.9)	11.5 (70.2)	9.2 (42.6)	11.9 (31.7)
Unvoiced Fricatives	*SS* (σ)	18.8 (10.5)	11.8 (10.4)	20.0 (13.0)	16.4 (28.2)	10.8 (11.0)
	MS (σ)	24.9 (16.0)	18.5 (12.9)	23.8 (15.2)	25.4 (18.6)	19.6 (12.2)
	DD (σ)	-12.3 (30.6)	-16.9 (22.6)	-13.9 (28.1)	-50.3 (40.9)	-22.3 (25.8)
Voiced Fricatives	*SS* (σ)	11.6 (20.6)	17.9 (40.4)	24.7 (40.7)	26.2 (46.0)	12.8 (16.5)
	MS (σ)	8.9 (16.3)	15.2 (35.1)	17.3 (25.0)	24.9 (41.6)	10.7 (15.7)
	DD (σ)	9.2 (24.5)	4.4 (28.8)	3.9 (42.8)	0.5 (43.8)	4.9 (23.5)
Nasal Consonants	*SS* (σ)	17.0 (29.8)	14.3 (10.5)	15.6 (17.1)	26.0 (23.2)	25.7 (36.0)
	MS (σ)	11.4 (26.3)	5.7 (6.2)	12.2 (13.4)	13.6 (13.7)	16.1 (35.8)
	DD (σ)	9.4 (24.4)	15.5 (23.4)	25.7 (34.4)	21.4 (32.0)	17.4 (26.2)
Oral Vowels	*SS* (σ)	12.9 (23.3)	12.8 (17.9)	16.7 (34.2)	22.4 (38.1)	16.6 (37.7)
	MS (σ)	11.8 (15.2)	16.7 (14.8)	17.0 (30.7)	20.1 (29.5)	14.5 (25.4)
	DD (σ)	1.5 (28.5)	5.8 (26.6)	4.4 (35.7)	15.6 (56.1)	3.1 (40.4)
Nasal Vowels	*SS* (σ)	9.0 (23.2)	7.5 (9.7)	14.2 (24.6)	22.6 (32.7)	13.6 (23.6)
	MS (σ)	9.2 (12.8)	6.6 (7.0)	13.2 (20.5)	19.4 (18.2)	10.6 (12.1)
	DD (σ)	8.3 (27.5)	9.4 (31.3)	-6.1 (52.6)	-3.4 (57.1)	7.2 (25.0)
All phones	*SS* (σ)	12.8 (22.2)	13.5 (20.6)	18.9 (34.3)	22.8 (37.7)	15.1 (29.8)
	MS (σ)	11.3 (16.1)	11.3 (14.8)	16.4 (26)	20.3 (29.9)	13.0 (21.5)
	DD (σ)	3.49 (28.2)	0.7 (30.2)	3.7 (39.8)	13.0 (50.6)	3.6 (34.2)

Figure 1: Distribution of Start Shift - *SS* (above) and Midpoint Shift - *MS* (below) values[3] given for the control group and per pathology. Each bin refers to a shift of 1 frame (10ms).



mal and dysarthric speech. Measures proposed in 3.2 are computed for each phone and averaged over the different pathologies present in the corpus (see section 3) and phonetic categories. The phonetic categories considered in this study are unvoiced plosives, voiced plosives, unvoiced fricatives, voiced fricatives, nasal consonants, oral and nasal vowels, providing 634, 293, 246, 1199, 228, 2115 and 223 phones respectively. Table 2 depicts the average values for *SS*, *MS* and *DD* measures and their standard deviations, all computed per pathology and phonetic category. Measures coming from the control speakers are also provided.

4.1. Control speakers vs. Pathologies

Regarding the overall set of phones, we observe that control speakers obtain measure values quite similar to those of LSD patients. On the other hand, there is much more variability among the other pathologies, illustrated by higher standard deviation values, notably for CA patients. Regarding measures per phonetic category, we can observed that :
- a better alignment is reached for control speakers

on the voiced plosives, compared with all the pathologies;
- alignments on the unvoiced plosives and voiced fricatives for both the control speakers and PD patients are quite similar, even *DD* are shorter for PD patients; similar behavior can be noted for the nasal vowels;
- values reached by LSD patients on the nasal consonants and vowels are better than those of the control speakers (with very low σ values). This can be explained by the hypernasality usually observed with mixed dysarthria. However, this behavior is not clearly observed here with ALS patients, also associated with a mixed dysarthria, for which much more variability is observed notably on the *DD* values;
- better values are reached on the unvoiced fricatives for LSD and PD patients compared with the control speakers while close values are observed for both other pathologies ;
- voiced plosives (fricatives) outperform unvoiced plosives (fricatives) for control speakers whereas the opposite is not systematically observed over pathologies. This could be caused by anomalies such as the voicing of voiceless consonants or the phenomena of spirantization (presence of low intensity friction noise in place of silence observed during "normal" plosive production) observed on pathologies.

4.2. Dysarthric speech

By comparing measure values over the different pathologies, we can note that :
- similar behavior is rather observed among ALS and PD patients on unvoiced plosives (with best values reached by ALS compared with the control speakers), among LSD and PD patients on voiced plosives and unvoiced fricatives ;
- ALS patients present the worst values on voiced plosives with the largest variability;
- a large variability is observed among LSD, ALS, and CA patients regarding the voiced fricatives, compared with PD patients, who reach the lowest values by far;
- LSD patients show the lowest values for the nasal consonants and vowels, the PD patients exhibiting the largest variability on the nasal consonants, especially for *SS* and *MD* values. Similar behavior is observed with the LSD patients on the oral vowels, the ALS, CA, and PD patients still exhibiting a very large variability ;
- CA patients are mostly associated with the worst values and the largest variability.

In order to complete these observations, figure 1 displays the *SS* (above) and *MS* (below) distributions[3] for all the phones in terms of frame numbers before (negative values) or after (positive values) the manual segmentation boundaries (0 bin refers to a shift between automatic and manual segmentation less than 1 frame i.e. less than 10ms). Considering that the range of acceptable values is from -2 to 2 frames (range usually used in phone segmentation evaluation), this distribution shows that patients suffering from PD or LSD present the smallest number of *SS* and *MS* values over this range. For instance, only 12.9% and 13.2% *MS* values with LSD and PD respectively are over this range compared to 21.7% and 28.5% with ALS and CA patients respectively.

4.3. Discussion

The observations reported in the previous sections highlight a quite large heterogeneity of behaviors depending on phonetic classes between pathologies, but also within a same pathology. Except the CA patients for which the automatic system encounters constant difficulties whatever the phonetic category it deals with, its behavior can be very different with the other pathologies depending on the phonetic categories. Concerning the CA patients, the constant difficulties encountered by the automatic system can be directly linked to the articulatory imprecision cluster defined in the Mayo classification [5] for the ataxic dysarthria, leading to the production of imprecise consonants and distorted vowels. Additionally, these patients, as expected, present the lowest perceptual speech rate among all pathologies, which can also explain the difficulties of the automatic system. The imprecision of consonants characterizes also the mixed dysarthria related to ALS and LSD pathologies as well as, with a more moderate level, the hypokinetic dysarthria related to PD pathology. However, this feature is not so clear when compared with the control speakers (best values reached by ALS patients on the unvoiced plosives, best values for LSD and PD patients on unvoiced fricatives, both compared to the control speakers). Moreover, observations done on ALS and LSD patients are not so consistent since LSD patients reach the best values on all the phones and present the highest Dysarthria Severity Degree values.

5. CONCLUSION

The observation of the behavior of a straightforward automatic speech processing tool when applied on dysarthric read speech produced by patients suffering from different pathologies has shown the very large variability that patients can demonstrate in their production of phonetic categories and their alterations. These observations may explain the difficulties of the automatic speech processing tools when they are faced to dysarthric speech. In future works, this study will be pursued on more patients, still on read speech but also on spontaneous speech.

6. REFERENCES

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

[2] Carmichael, J. 2007. *Introducing objective acoustic metrics for the Frenchay Dysarthria Assessment procedure*. Ph.d. dissertation, university of sheffield.

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

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

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

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

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

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

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

[10] Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., Gravier, G. September 2005. Ester phase ii evaluation campaign for the rich transcription of french broadcast news. *Proceedings of Interspeech'05* 1149–1152.

[11] Green, J. R., Yunusova, Y., Kuruvilla, M. S., Wang, J., Pattee, G. L., Synhorsti, L., Zinman, L., Berry, J. D. 2013. Bulbar and speech motor assessment in als: Challenges and future directions. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 14(7–8), 494–500.

[12] Green, P., Carmichael, J., Hatzus, A., Enderby, P., Hawley, M., Parker, M. 2003. Automatic speech recognition with sparse training data for dysarthric speakers. *Proceedings of Interspeech'03* Geneva, Switzerland. 1189–1192.

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

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

[15] Kim, M., Kim, H. 2012. Automatic assessment of dysarthric speech intelligibility based on selected phonetic quality features. In: *Computers Helping People with Special Needs* volume 7383 of *Lecture Notes in Computer Science*. 447–450.

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

[17] McAuliffe, M. J., Ward, E. C., Murdoch, B. E. 2006. Speech production in parkinson's disease: I. an electropalatographic investigation of tongue-palate contact patterns. *Clinical linguistics & phonetics* 20(1), 1–18.

[18] Middag, C., Martens, J.-P., Nuffelen, G. V., Bodt, M. D. 2009. Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Applied Signal Processing* 2009(1).

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

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

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

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

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

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

[25] Whitfield, J. A., Goberman, A. M. 2014. Articulatory-acoustic vowel space: Application to clear speech in individuals with parkinson's disease. *Journal of communication disorders* 51, 19–28.

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

[1] 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.

[2] compared with more complex technologies like automatic speech recognition

[3] Here the *SS* and *MS* measures are given without considering the absolute value.