Multivariate Analysis of Vocal Fatigue in Continuous Reading

Marie-José Caraty [1], *Claude Montacié* [2]

[1] Paris Descartes University - LIPADE, France
[2] Paris Sorbonne University - STIH, France

Marie-Jose.Caraty@ParisDescartes.fr, Claude.Montacie@paris-sorbonne.fr

Abstract

We present an experimental paradigm to measure changes in characteristics of speech under vocal fatigue. As speech material, we have chosen a vocal load (3 hours) and a cognitive process (aloud continuous reading) that can induce some fatigue of the reader. The fatigue is verified using an analysis of reading errors and of disfluencies. A multivariate analysis based on Wilks' lambda test, on 169,042 occurrences of phonemes, allows to analyze spectral and prosodic changes on each phonetic class. The results upon six readers show that the nasals (vowels and consonants) are the phonemes the most discriminative in vocal fatigue.

Index Terms: vocal fatigue, non-linguistic cues, phonetic classes, multivariate analysis

1. Introduction

Our study on vocal fatigue uses an experimental paradigm for measuring changes in vocal observations under induced condition of vocal fatigue. Many articles in the experimental literature on vocal fatigue describe conditions such as sleep deprivation and/or sustained/overload task [1-5]. The nature of the fatigue which is induced by such processes is physiologic and/or neurologic. The task we have chosen for this study is the aloud continuous reading during about 3 hours which is a quite long duration for reading.

In psychology, many studies on reading investigate the role played by the speech production during reading. From visual to meaning, several hypotheses exist of speech recoding in the reading which are always debated [6]. This recoding consists in the transformation from printed words to a speech-based code : an intermediary stage before understanding. The nature of the speech units involved in the cognitive process is a part of the debate (articulatory, acoustic, auditory imagery or a more abstract code). Reading is a skill that involves a complex cognitive processing [6-7]. For a trained or a professional reader, the skill consists in reading and in interpreting fluently the content with a prosody suitable to the emotional intention of the author.

Many studies on vocal fatigue in speech pathology [8, 9] investigate the effect of vocal load on voice disorders. The symptoms of vocal fatigue are various and explained by the physiologic mechanisms of vocal production. Among the important contributors to vocal production and among all the muscles involved in the vocalization, the vocal folds are the most studied in the works related to the neuromuscular fatigue. In prolonged vocal use, the vocal folds seem to sustain an increasing viscosity and a decreasing in blood circulation, these are probable factors inducing some voice disorders [10]. An overloading of vocal use induces a vocal fatigue as it is suggested/shown in various studies [8-10], even if the fibers of vocal folds are highly resistant.

Publications related to vocal fatigue induced by a task are numerous [8-13] and their results are difficult to compare. In related work experiences on reading, the reading task is very variable according to –the read content (vowels, isolated sentence, text), –the production constraint on pitch and intensity level, –the background noise level, –the duration of reading (from 20 to 150 minutes), –the number of readers (from 1 to 50). In related work, a high number of prosodic and/or spectral parameters have been investigated and the results often conflicting [10]. For the discussion, we will give our findings on spectral and prosodic parameters.

2. Speech material

For our investigation on vocal fatigue, we have chosen speech material with a vocal loading (3 hours) and a cognitive process (loud continuous reading) that can induce fatigue of the speaker. Six readers are considered in the experiments for a relative speaker independency of the results.

2.1. Speech corpora

Speech corpora come from recordings of Direct 8 TV program « Voyage au bout de la nuit ». The program consists in continuous reading of French literary works by young actors. The program duration is around 3 hours, shared into six periods giving six speech segments (*s1* to *s6*) of 30 minutes of continuous reading. The six periods are consecutively recorded with a short pause between two periods. The program is then composed of advertising followed by a brief summary of the previous period and the reading of the next period. Six program recordings [14] are used for the experiments, with six different readers (*R1* to *R6*) composed of five actresses (*R1* to *R5*) and one actor (*R6*).

2.2. Speech and text corpora processing

For analytic experiments, we have listened to the segments *s1* and *s6* for each speaker and corrected the original text according to the text uttered by the reader.

The D-DAL speech transcription system [15] is used to align the speech signal with the corrected text. Alternative phonetic transcriptions of word allow to take into account variations in pronunciation. The time code in speech (in cs) of the beginning and the end for each word and each phoneme of the word are computed. The inputs of the alignment system are the whole duration speech segment and the corresponding text.

Six hours of speech (one hour per reader) are so automatically aligned, to constitute a database of 54,520 occurrences of words of the 7,246 different words and their 169,042 corresponding occurrences of phonemes. The quality of the time alignment has been assessed by random listening of 600 words so aligned. No error appeared at the word level. The table 1 gives the percentage of occurrences of the phonemes (SAMPA code) in the database.

Ph.	a	a~	b	d	e	E	e~	@	2
Occ (%)	8.1	3.4	1.2	4.8	5.2	5.6	1.4	3.5	0.6
Ph.	9	f	g	H	i	j	J	k	l
Occ (%)	0.7	1.3	0.6	0.6	5.1	1.7	0.2	3.6	6.7
Ph.	m	n	o	O	o~	p	R	s	S
Occ (%)	3.5	2.8	1.3	2.0	1.9	3.3	8.1	5.6	0.7
Ph.	t	u	v	w	y	z	Z		
Occ (%)	5.6	2.3	2.7	1.0	2.1	1.4	1.4		

Table 1. *Occurrence percentage (**Occ**) of phonemes (**Ph**) on the 169,042 occurrences in the database.*

2.3. Spectral and prosodic features extraction

The acoustic features were extracted for each phoneme from the speech corpus using the Praat software [16]. For each phoneme, a vector of 12 Mel frequency cepstral parameters is computed at the center of the phoneme using a 15 ms Hamming windowed frame. Seven other parameters are computed for voiced phonemes (excluding *p*, *t*, *k*, *f*, *s* and *S*) : the fundamental frequency (*f0*), the duration (*d*) the jitter value (*jt*), the shimmer value (*sh*) and the first three formants (*f1*, *f2*, *f3*). The shimmer value is the average absolute difference between the amplitudes of consecutives periods, divided by the average amplitude. The jitter value is the average absolute difference between consecutives periods divided by the average period. The fundamental frequency and formants are estimated at the center of the phoneme. The shimmer and jitter values are estimated on the duration of the phoneme.

3. Analysis of reading errors and disfluencies

To estimate the fatigue level of readers, the differences between the spoken text and the original text are analyzed. Two types of errors are made by the reader : reading errors and disfluencies. The reading errors mainly observed are i) the utterance of a word for another, generally close phonetically or ii) the omission of a part of the text coming from a confusion of lines. Disfluencies occur in speech reading as in spontaneous speech [17] even if the cognitive process in reading is fundamentally different. The disfluencies observed are mainly i) word repetitions, ii) fragments of words with the various types of restarts or interruption. Only the filled pauses are not so frequently observed than in spontaneous speech.

For a reliable estimator of reading errors and disfluencies : the deletion, insertion and substitution errors of phonetic segments are considered. This error in deletion, insertion and substitution is computed with the Wagner and Fischer algorithm between the phonetic transcriptions of the spoken text and of the original text. The table 2 gives the error rates per minute computed from the 3 components of error (deletion *D*, insertion *I* and substitution *S*) over the segments *s1* and *s6* (of duration *d* in minutes) for each reader. The average (*Av*) over all readers is given for each measure on *s1* and *s6*.

At the beginning of *s1*, the reader is supposed to be in a state of non-vocal fatigue whereas at the beginning of *s6* the reader yet sustained a vocal loading of 2h30 in continuous reading. The results of the total error rate shows clearly, for all readers, reading errors and disfluencies are indicators of cognitive fatigue. For all readers, the error rate increases from *s1* to *s6*. For *R3*, the increasing reaches a factor 3 from *s1* to *s6* and the average over the readers is 1.6 time higher on *s6* than on *s1*.

R.	seg.	d	D	I	S	Error
R1	s. 1	30.9	1.4	1.8	1.6	4.7
	s. 6	23.7	1	2.3	1.6	4.9
R2	s. 1	28.5	0.4	1.2	0.4	1.9
	s. 6	28	0.4	1.7	0.5	2.5
R3	s. 1	28.5	0.1	1	0.2	1.3
	s. 6	28.9	1.6	2.1	0.6	4.3
R4	s. 1	28.4	0.3	0.7	0.2	1.2
	s. 6	28.1	0.6	1.6	0.5	2.7
R5	s. 1	28.3	0.9	2.2	1.4	4.6
	s. 6	28.7	2.2	2.9	1.7	6.9
R6	s. 1	27.7	0.7	1.2	0.7	2.6
	s. 6	29.6	0.5	2.9	1.2	4.5
Av.	s. 1	28.7	0.6	1.4	0.8	2.7
	s. 6	27.8	1.1	2.3	1.0	4.3

Table 2. *Error rates per minute (**D**eletion, **I**nsertion, **S**ubstitution and total **Error**) according to the reader (**R1** to **R6**), on the segments (**s1** and **s6)** of given **d**uration (minutes) and averages (**Av**) over readers.*

For the following experiments, we assume that during the speech segment *s6* the reader is under vocal fatigue.

4. Performance of discriminant analysis on spectral and prosodic parameters

The discriminant analysis was used to detect the most discriminating phonemes showing differences in feature vectors between the segments *s1* and *s6*. This analysis is based on multivariate statistics of the two sets of feature vectors for a given reader, the first selected on *s1*, the second on *s6*.

4.1. Multivariate analysis of variance

The Multivariate Analysis of Variance (MANOVA) is the multivariate analog of the Analysis of Variance (ANOVA) used in univariate statistics. It is based on the decomposition of the total sum of squares and cross products matrix (*T*) into two matrices : the Hypothesis sum of squares and cross products matrix (*H*) and the Error sum of squares and cross products matrix (*E*) [18]. The maximum discrimination is obtained when *T* is close to *H* and there is no discrimination when *T* is close to *E*. Four methods have been developed to measure discrimination : the Wilks' lambda, the Hotelling's Trace, the Pillai's Trace and the Roy's largest root [19]. We used the first two methods for this study. As the results of these two methods were consistent, we present the results only with the Wilks' lambda method. This Wilks' lambda varies from zero (*T* close to *H*) to one (*T* close to *E*). To assess the quality of measurement, a probability value (p-value) is associated to each Wilks' lambda.

4.2. Discriminant analysis on spectral parameters

The search on the most discriminating phonemes consists in computing the Wilks' lambda for all pairs of readers and phonemes. This value is estimated on the feature vectors of *s1* and *s6* using the R software [20]. To calibrate theses values, a global Wilks' lambda is also estimated on all feature vectors of a reader. To ensure quality results, all the Wilks' lambda of a p-value below the usual threshold of 0,001 have not been taken into account. On the 210 Wilks' lambda computed (6 readers x 34 phonemes + 10 global), 189 have a p-value below the threshold. We decided to study only the 22 phonemes for which a significant Wilks' lambda was calculated for each reader. The rejected phonemes are *b*, *9*, *2*, *f*, *g*, *H*, *j*, *J*, *p*, *S*, *w* and *Z*. There are multiple reasons for a p-value too high (low number of occurrences as in the case of phoneme *J*, inherent variability as in the case of phoneme *p*). The results for the 22 phonemes and all classes (*Gl*), sorted by the average of the Wilks’ lambda over all readers, are presented in the table 3.

Ph.	R1	R2	R3	R4	R5	R6	Average
a~	.77	.89	.64	.47	.39	.54	.61
o~	.71	.85	.74	.52	.37	.54	.62
m	.68	.81	.72	.63	.39	.51	.62
n	.66	.80	.72	.59	.37	.65	.63
e~	.79	.84	.72	.50	.38	.74	.66
@	.79	.85	.78	.61	.37	.82	.70
E	.83	.96	.80	.54	.43	.86	.74
O	.74	.95	.86	.57	.46	.91	.75
e	.87	.92	.86	.57	.42	.85	.75
a	.82	.94	.82	.59	.46	.89	.76
v	.84	.96	.76	.52	.55	.93	.76
o	.73	.92	.85	.67	.56	.89	.77
z	.82	.92	.87	.57	.70	.90	.79
i	.87	.94	.82	.65	.70	.85	.81
y	.85	.90	.77	.70	.72	.89	.81
u	.89	.87	.83	.75	.70	.83	.81
s	.96	.97	.77	.72	.57	.94	.82
l	.85	.93	.92	.70	.72	.86	.83
R	.89	.97	.84	.62	.74	.92	.83
d	.85	.97	.87	.72	.73	.91	.84
t	.95	.98	.92	.78	.86	.92	.90
k	.95	.97	.95	.81	.91	.95	.92
Gl	.94	.98	.91	.78	.83	.94	.90

Table *3. Discrimination using 12 MFCC parameters Wilks’ lambda, per phonetic class* **(Ph)** *and reader* **(R1** *to* **R6)** *ranked by the* **Average** *over readers.*

This table gives several interesting results. The first result is that the most discriminating phonemes are the nasals (vowels and consonants) as shown by the 5 first ranks by the average Wilks’ lambda ranking. Summing, over the readers, the occurrences of the phonemes inside the 3 first ranks, we find the phonemes the most stables in a reader-independent context are in order {*m*, *n*, *o~*}, {*a~*}, {*e~*} and {*@*}. The second result is the high value of global Wilks' lambda (*Gl*). This indicates that it is difficult to measure general spectral changes due to an overall prolonged vocal effort.

4.3. Discriminant analysis on prosodic and spectral parameters

Many studies on vocal fatigue have investigated various prosodic parameters such as the duration of phonemes, the fundamental frequency, the jitter and the shimmer. It is interesting to verify whether the addition of these parameters and the first three formants, could improve the discrimination capability. For this experiment, we selected the 113,322 phonetic segments of voiced phonemes (excluding *p*, *t*, *k*, *f*, *s* and *S*) where these parameters can be estimated by the software Praat. The table 4 gives the percentage of occurrence of the phonemes used for this experiment.

On the 168 Wilks' lambda computed (6 readers x 28 phonemes remaining), 151 have a p-value below the threshold (0.001). We decided to study only the 18 phonemes for which a significant Wilks' lambda was calculated for each reader. The rejected phonemes are *b*, *9*, *2*, *g*, *o*, *H*, *j*, *J*, *w* and *Z*.

Ph.	a	a~	b	d	e	E	e~
Occ (%)	11	5.0	1.7	6.0	7.0	7.8	2.1
Ph.	@	2	9	g	H	i	j
Occ (%)	4.9	0.8	0.9	0.6	0.5	6.1	1.7
Ph.	J	l	m	n	o	O	o~
Occ (%)	0.4	8.2	4.9	4.0	1.8	2.8	2.7
Ph.	R	u	v	w	y	z	Z
Occ (%)	6.8	2.9	3.4	1.3	2.2	1.5	1.1

Table 4. *Occurrence percentage* **(Occ)** *of phonemes* **(Ph)** *on the 113,322 occurrences in the experiment.*

The results for the 18 phonemes, sorted by the average of the Wilks’ lambda over all readers, are presented in the table 5.

Ph.	R1	R2	R3	R4	R5	R6	Average
m	.64	.75	.65	.51	.35	.48	.57
o~	.69	.83	.65	.48	.33	.48	.58
n	.59	.79	.64	.51	.34	.58	.58
a~	.73	.87	.62	.43	.36	.46	.58
e~	.74	.82	.68	.43	.34	.62	.60
@	.77	.79	.72	.55	.36	.76	.66
E	.79	.94	.77	.50	.39	.76	.69
O	.70	.92	.77	.53	.42	.83	.70
z	.68	.85	.78	.44	.64	.81	.70
v	.80	.94	.70	.44	.50	.85	.70
e	.85	.90	.79	.52	.38	.80	.70
a	.80	.92	.80	.54	.42	.84	.72
y	.82	.89	.68	.61	.55	.80	.73
u	.82	.84	.76	.67	.53	.74	.73
i	.81	.93	.74	.60	.61	.76	.74
l	.79	.91	.86	.62	.57	.80	.76
R	.82	.95	.82	.59	.58	.88	.77
d	.84	.94	.79	.66	.64	.86	.79

Table 5. *Discrimination using 19 parameters Wilks’ lambda, per phonetic class* **(Ph)** *and reader* **(R1** *to* **R6)***, ranked by the* **Average** *over readers*

The table shows that the addition of the prosodic parameters increases significantly the discrimination performance for all phonemes. The most discriminating phonemes remain the nasal phonemes (vowels and consonants). We may also notice that the schwa (@) is still in 6th rank. As in the previous experiment, in a reader-dependent context the discrimination the most stable is for {*m*, *n*, *o~*}, {*a~*} and {*@*}.

4.4. Parameters selection

The goal in this experiment is to reduce the number of parameters by removing the least discriminating. The selection parameters algorithm is as follows : let be, *E* a set of parameters, and *w(E)* the average of Wilks' lambda computed with the parameters of *E* over all 6 readers and 18 phonemes. At the beginning, let be *E1*, the set of all 19 parameters.

At the step *i*, let be P_i is the parameter of the set E_i defined by the equation 1, and E_{i+1} the set of parameters at step *i+1*, E_{i+1} is defined by the equation 2 :

$$P_i = \underset{\{P_j\} \in E_i}{\text{Argmin}} \left\{ w\left(E_i - P_j\right) \right\} \tag{1}$$

$$E_{i+1} = E_i - P_i \tag{2}$$

The results of the algorithm are presented in the table 6 with the parameter P_i rejected at the step *i*, and the value $w(E_i)$.

Step i	1	2	3	4	5	6	7
$w(E_i)$	0.68	0.68	0.69	0.69	0.69	0.70	0.70
P_i	f3	f1	f2	jt	d	sh	c7
Step i	8	9	10	11	12	13	14
$w(E_i)$	0.71	0.72	0.73	0.74	0.76	0.78	0.80
P_i	c11	c9	c12	f0	c6	c4	c5
Step i	15	16	17	18	19		
$w(E_i)$	0.82	0.86	0.87	0.90	0.94		
P_i	c10	c3	c8	c2	c1		

Table 6. *Parameter rejected (**P_i**) and average of Wilks' lambda (**$w(E_i)$**) according to the step **i** of the selection algorithm of parameters.*

As shown in the table 6, the least discriminating parameters are formants and prosodic parameters except the fundamental frequency. The most discriminating parameters are the first cepstral coefficients. We also note that the value of the function *w* corresponding to the 12 MFCC (0.72) is similar to the value of the function *w* corresponding to the 11 most discriminating parameters (*c1*, *c2*, *c8*, *c3*, *c10*, *c5*, *c4*, *c6* and *f0*, *c12* and *c9*).

5. Discussion and conclusion

In our study, the salient and unexpected result is that, over all the readers, the nasals (vowels and consonants) offer better discrimination than the other classes of phonemes (except the rejected phonemes on which we can't conclude). This could be explained by a significant effort required to produce the nasals. Indeed, access to the nasal cavity is controlled by lowering (open access) or raising (closed access) the velum/soft palate. It is possible that a prolonged vocal load disrupts this control. Furthermore, as assumed for vocal folds, the velum is also affected by the same phenomenon of dehydration by the airflow inducing a possible increasing in viscosity. Of course, these assumptions have to be verified with specific studies on speech production.

Previous studies in vocal fatigue are varied according to the task which induces the vocal load and the results are often conflicting [10]. In our study of the vocal fatigue upon six readers, the spectral parameters are shown to be more discriminating than prosodic parameters, only f0 seems to have a performance in discrimination ; the shimmer, the duration and the jitter are far from this performance. At last, previous studies have looked for a general change of speech features. Our results on global statistics versus statistics per phonetic class show these changes are different depending on the class. These findings suggest the need of a phoneme spotting system for the design of a vocal fatigue detector.

The perspectives for future works are : i) test more speakers and on other languages to check if the nasals phonemes are robust in the vocal fatigue discrimination, ii) test whether fatigue induced by sleep deprivation causes the same effects, iii) detect spoken sentences with significant vocal fatigue, iv) correlate the results of the detection with the positions of disfluencies and/or reading errors, v) study the feasibility of a vocal fatigue detector.

6. References

[1] Brenner, M., Doherty, E. T. and Shipp, T., "Speech Measures Indicating Workload Demand", Aviation, Space, and Environmental Medicine, 65, pp. 21-26, 1994.

[2] Bard, E. G., Sotillo, C., Anderson, A. H., Thompson, H. S., and Taylor, M. M., "The DCIEM Map Task Corpus : Spontaneous Dialogue under Sleep Deprivation and Drug Treatment", Speech Communication, 2, pp. 71–84, 1996.

[3] Whitmore, J. and Fisher, S., "Speech During Sustained Operations", Speech Communication, 2, pp. 55–70, 1996.

[4] Nwe, T. L., Li, H. and Dong, M., "Analysis and Detection of Speech under Sleep Deprivation", Interspeech, pp.17-21 ,2006.

[5] Pilcher, J. J., McClelland, L. E., Moore, D. D., Haarmann, H., Baron, J., Wallsten, T. S. and McCubbin, J. A., "Language Performance Under Sustained Work and Sleep Deprivation Conditions", Aviation, Space, and Environmental Medicine, Vol. 78, Supplement 1, pp. B25-B38, 2007.

[6] Kleiman, G. M., "Speech Recoding in Reading", Journal of Verbal Learning and Verbal Behavior, 14, pp. 323-339, 1975.

[7] Rayner, K. and Clifton, C., "Language processing in reading and speech perception is fast and incremental : Implications for event-related potential research", Biological Psychology 80, pp. 4–9, 2009.

[8] Solomon, N. P., Glaze, L. E., Arnold, R. R. and van Mersbergen, M., "Effects of a Vocally Fatiguing Task and Systemic Hydration on Men's Voices", Journal of Voice, Vol. 17, No. 1, pp. 31-45, 2003.

[9] Laukkanen, A. M., Ilomäki, I., Leppänen, K. and Vilkman, E., "Acoustic Measures and Self-reports of Vocal Fatigue by Female Teachers", Journal of Voice, Vol. 22, No. 3, pp. 283-289, 2008.

[10] Welham, N. V. and Maclagan M. A, "Vocal Fatigue: Current Knowledge and Future Directions", Journal of voice, Vol. 17, No. 1, pp. 21-30, 2003.

[11] Pausewang Gelfer, M., Andrews, M. L. and Schmidt, C. P., "Effects of prolonged loud reading on selected measures of vocal function in trained and untrained singers", Journal of Voice, Vol. 5, No. 2, pp. 158-167, 1991.

[12] Greeley, H. P., Berg, J., Friets, E., Wilson, J., Greenough, G., Picone, J., Whitmore, J. and Nesthus, T., "Fatigue estimation using voice analysis", Behavior Research Methods, 39 (3), 610–619, 2007.

[13] de Bruijn, C. and Whiteside, S., "Use of Speech Recognition and Voice Fatigue : Measures of F0 and Spectral Slope", ICPHS, pp. 2061-2064, 2007.

[14] Direct 8 TV programme « Voyage au bout de la nuit »
Recording April 16, 2009. Reader (R1) : A. Blanquart, Extract from "Germinal", Emile Zola (1885).
Recording July 18, 2009. Reader (R2) : C. Beauxis, Extract from "Les trois mousquetaires", Alexandre Dumas (1844).
Recording July 21, 2009. Reader (R3) : E. Palle, Extract from "Les trois mousquetaires", Alexandre Dumas (1844).
Recording September 16, 2009. Reader (R4) : L. Pasteau, Extract from "Sans Famille", Hector Mallot (1878).
Recording September 23, 2009. Reader (R5) : V. Hilssone, Extract from "Le petit chose", Alphonse Daudet (1868).
Recording July 16, 2009. Reader 6 : L. Cécilio, Extract from "Les trois mousquetaires", Alexandre Dumas (1844).

[15] Montacié, C. and Caraty, M.-J. "Sound Channel Video Indexing,", ESCA, Eurospeech 97, pp. 2359-2362, 1997.

[16] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer [Computer program]", Version 5.1.31, retrieved 4 April 2010 from http://www.praat.org/, 2010.

[17] Stouten, F., Duchateau, J., Martens, J.-P. and Wambacq, P., "Coping with disfluencies in spontaneous speech recognition : Acoustic detection and linguistic context manipulation", Speech Communication, 48, pp. 1590-1606, 2006.

[18] Mathew, T., "MANOVA in the multivariate components of variance model", Journal of Multivariate Analysis, Vol. 29, Issue 1, pp. 30-38, 1989.

[19] Thompson, G. L., "A unified approach to rank tests for multivariate and repeated measures designs", J. Amer. Statist. Assoc., 86, pp. 410-419, 1991.

[20] R Development Core Team, "R: A Language and Environment for Statistical Computing", R Foundation for Statistical Computing, Vienna, Austria, See http://www.R-project.org, 2008.