Exploring Acoustic Parameters of Patients Treated for Oral Cancer to Assess the Severity of Pathological Speech

Etienne Sicard

Exploring Acoustic Parameters of Patients Treated for Oral Cancer to Assess the Severity of Pathological Speech

Etienne Sicard[1], *Julie Mauclair*[2, 3], *Jérôme Farinas*[3], *Julien Pinquier*[3]

[1]INSA, University of Toulouse, France
[2]Paris Descartes University, Paris, France
[3]IRIT, University of Toulouse, UPS, Toulouse, France

etienne.sicard@insa-toulouse.fr, {julie.mauclair,jerome.farinas,pinquier}@irit.fr

Abstract

As part of the Carcinologic Speech Severity Index project (C2SI), the voices of 35 patients who were treated for oral and pharyngeal cancer were studied using different acoustic parameters, coupled with automated speech recognition. The study revealed, by comparison with a control group, significant harmonic poverty on sustained /a/ as well as on read text. A spectral degradation of vowels and consonants, in particular in the areas of formants F2 to F4 was also observed. No significant difference in the characteristics of pitch histogram was found. The consonant frequency distance /s/m/ computed from average spectrum is correlated with patient intelligibility.

Index Terms: Cancer, voice, acoustic indicators, LTASS, phonetic segmentation, harmonic poverty

1. Introduction

The decreasing mortality of Head and Neck Cancers highlights the importance to reduce the impact on Quality of Life. But, the usual tools for assessing QoL are not relevant for measuring the impact of the treatment on the main functions involved by the sequellae. Validated tools for measuring the functional outcomes of carcinologic treatment are missing, in particular for speech disorders. Some assessments are available for voice disorders in laryngeal cancer but there are based on very poor tools for oral and pharyngeal cancers involving more the articulation of speech than voice. In the C2SI project [8], we propose to develop a severity index of speech disorders describing the outcomes of therapeutic protocols completing the survival rates. Intelligibility of speech is the usual way to quantify the severity of neurologic speech disorders. But this measure is not valid in clinical practice because of several difficulties as the familiarity effect of this kind of speech and the poor inter-judge reproducibility. Moreover, the transcription intelligibility scores do not accurately reflect listener comprehension.

It is acknowledged that an unbiased and objective assessment of the communication deficiency caused by a speech disorder calls for automatic speech processing tools. The principle is to perform an audio recording of the patient speaking and to compute the intelligibility of the utterances produced in the aim to obtain a score. Middag [1] presented a method that predicts running speech intelligibility in a way that is robust against changes in the text and against differences in the accent of the Dutch speakers applicable to patients treated for HNC. In this project [8] the main objective is to demonstrate that the carcinologic speech severity index (C2SI), obtained by an automatic speech processing tool, produces equivalent or superior outcomes than a score of speech intelligibility obtained by human listeners, in terms of quality of life foreseeing the speech handicap, after the treatment of oral and/or pharyngeal cancer. The analysis of speech and voice production for patients with cancer of the head and neck has been the subject of intensive research [2-6], both from the viewpoint of voice characteristics in relation with chemo-radiotherapy [4], laser microsurgery [5], as well as the construction of automatic tools to perform speech, phonation and voice intelligibility evaluation [6], which may guide speech-language pathologists towards efficient voice therapy [7].

In order to determine a predictive model of intelligibility on the basis of automatic speech analysis, we studied in this paper different approaches in order to discriminate pathological and normal voices.

2. Experimental Protocol

In this study, voices of 35 patients treated for oral and pharyngeal cancer by surgery, radiotherapy and/or chemotherapy was recorded at least 6 months after the end of treatment. The cancer is classified as stage T2 to T4 according to the TNM classification [9]. A group of healthy speakers was also established to serve as reference for this study. They are part of the first corpus of C2SI project [8] that was recorded at least 6 months after the end of treatment. Figure 1 gives a representation of the age of patients (vertical axis) versus the fundamental frequency of the patient's voice. It can be noticed that most patients are within the age group 55-75 years old, with a majority of male speakers (63%). The recording of the voice was conducted according to an identical protocol, hardware and software. The recorded sound includes a sustained /a/, the reading of a text extracted from the novel "Mr. Seguin's Goat", by Alphonse Daudet [10], including all vowels and consonants of the French language.

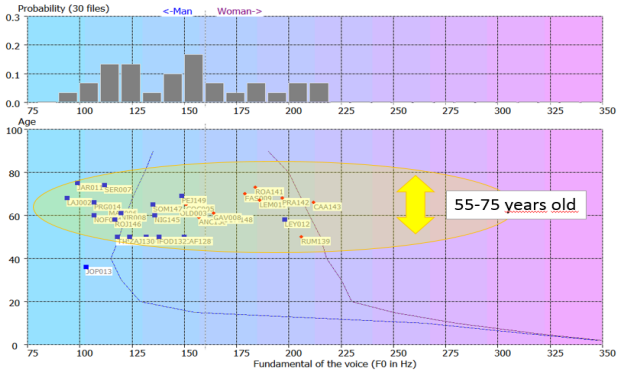

Figure 1: *Age vs. F0 of 30 patients analyzed within the C2SI project.*

The reproduction of a sound model of 50 pseudo-random words, and various other tasks such as description of scenes are also recorded. The sound files from the recordings were equalized, cleaned in order to keep the essential informative part produced by the patients. Various medical information such as age, sex, type of tumor, treatment, intelligibility and voice handicap scores have also been made available by the medical team: Speech Handicap Index [13], Phonation Handicap Index [14], as well as subjective evaluation of speech intelligibility from 4 voice pathologists and therapists. The resulting intelligibility (Figure 2) score was obtained, ranging from 0 (low) to 10 (high intelligibility), which serves as medical reference for correlations. The resulting intelligibility score uses a [0..10] scale, where 0 corresponds to very poor intelligibility (pathological voices) and 10 corresponds to perfect intelligibility (normal voice). The normal speaker's scores ranger from 8 to 10. The patient's score are spread almost over the entire scale [1..9.5], as shown in Figure 2.

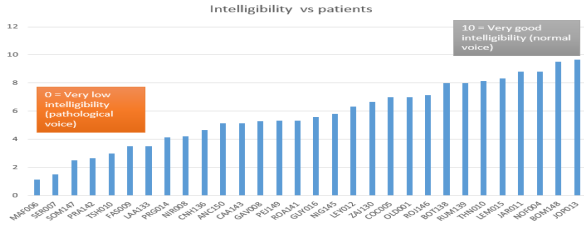

Figure 2: *Evaluation of the patient intelligibility based on a combination of VHI, PHI and medical team subjective evaluation of speech.*

3. Analysis

We have performed three studies in order to extract salient features which might differentiate normal speakers and patients. In 3.1, we wanted to know if the voice alteration in a sustained /a/ can be correlated with intelligibility. In 3.2, the analysis was conducted on pitch fluctuations and long-term spectrum. In 3.3 was conducted on an automatic phonetic segmentation of the read text.

3.1. Voice alteration: sustained /a/

A comparison between normal and pathological /a/ is reported in Figure 3, with the narrow-band spectrogram of the sound on the top (Frequency 0-2500 Hz in Y axis, time in X axis), and the low-frequency/high-frequency profile of the energy on the bottom.

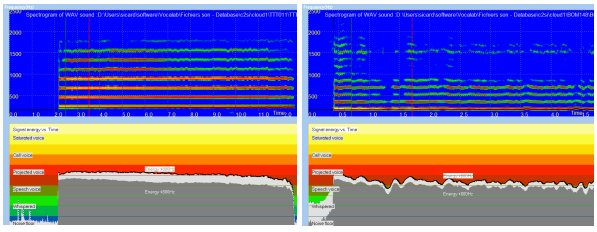

Figure 3: *Spectrogram and energy profile of the sustained /a/, normal (left) and pathological (right).*

The energy profile of the normal voice (Figure 3 left) is stable during the whole sample, while severe amplitude instability (Shimmer) is revealed for the patient (Figure 3 right). We clearly observe interruptions in the harmonic contents of the /a/, 5 voice pathology indicators [11] validated by French Speech and Language Pathologists have been evaluated on sustained /a/, each assessing specific characteristics of the voice:

- *Attack alteration*: evaluates the first 300 ms in terms of pitch instability, amplitude ramp-up and noise/harmonics level, combined into a single indicator.
- *Pitch Instability*: combines mid, long and very-long term Jitter.
- *Amplitude Instability*: combines mid, long and very-long term Shimmer.
- *Noise/Signal ratio*: combines conventional HNR measurement with hoarseness detection.
- *Harmonic poverty*.

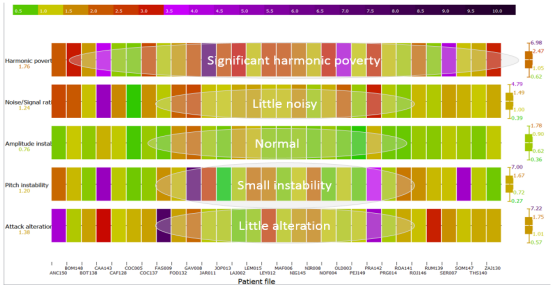

Figure 4: *5 voice alteration indicators of VOCALAB [10] on 30 patient recordings of sustained /a/. The green color indicates a normal value, lower than 1.0.*

Figure 5: *5 voice alteration indicators of VOCALAB [10] on 13 sustained /a/ from normal patients. All averages are significantly below the threshold.*

The normal/pathological limit is 1.0 for the 5 indicators. As seen in Figure 4, a significant alteration of the harmonic poverty is observed, with an average of 1.76, significantly higher than the threshold. In contrast, the normal group averages are below 1.0 for all indicators (Figure 5).

A comparison of the typical profile of the /a/ spectrum between normal speakers and patients reconfirms the previous observations: the average amplitude profile of the sustained /a/ from 60 normal speakers differs from the C2SI patients mostly in the F3-F4 formant zones, where a harmonic depletion around -15 dB is observed, which is strongly correlated to the "Harmonic Poverty" indicator (Figure 6).

The same observation can be made on long-term average speech spectrum (LTASS [15]) based on the reading of a text, where again the patient group exhibits a significant energy reduction in the high-frequency formants. The lack of resonance may be the consequence of cancer treatment and surgery, which affects the oral cavities, tongue mobility and inner resonators close to the vocal folds.

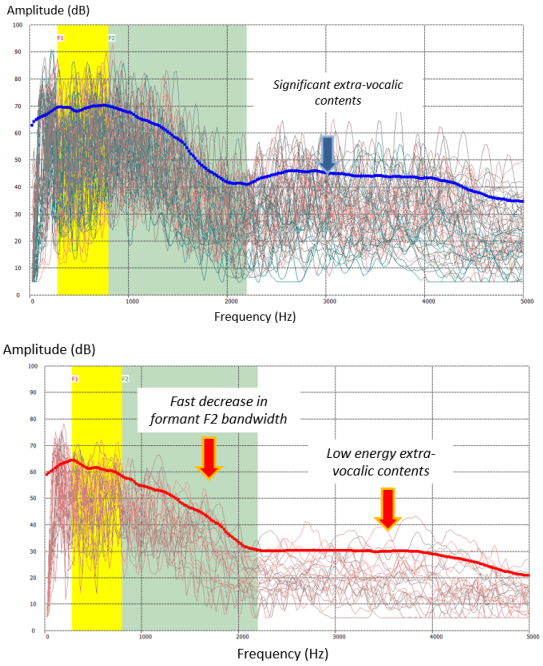

Figure 6: *Average Speech Spectrum: normal sustained /a/ on 60 different speaker (top) vs. 30 sustained /a/ of patient voices after cancer treatment (patients).*

3.2. Analysis of pitch fluctuations and long-term spectrum

The analysis of the text reading was carried out using two types of tools: *VOCALAB* [11], a tool used for evaluation & speech therapy, and the *OpenSmile* library [12].
The histogram of the fundamental frequency extracted from 20-ms portions of speech enables to compute the pitch variations and its associated parameters: standard deviation, kurtosis, skewness. Surprisingly, the comparison between the patient and control group did not reveal any significant difference in the pitch variation (mean/SD 5.5/1.4 notes for patients, 6.1/1.1 for normal speakers), which corresponds to normal standards. The In other words, patients recovering from cancer treatment compensate the significant harmonic poverty noticed in previous analysis by large pitch variations, to increase intelligibility. A very different interpretation could be that the cancer treatment may lead in some case to excess pitch fluctuations in sustained phonation, as reported for some variants of Parkinson's disease [16].

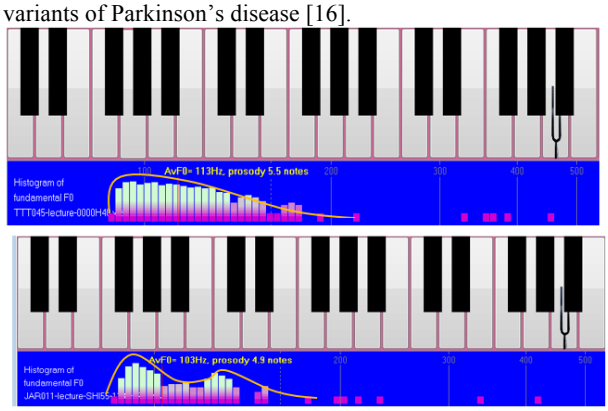

Figure 7: E*xamples of non-Gaussian distributions of F0: probability density shifted left (a) and low probability density in the middle (b).*

Furthermore, no significant correlation (0.24) was found between the degree of severity and the F0 distribution moments. Only in some rare cases, abnormal skewness or kurtosis was observed, such as reported in Figure 7, affecting both the patient and normal groups.

3.3. Phoneme analysis

An automatic alignment has been performed on the beginning of reading passage [10] of the corpus of C2SI project [8]. The alignment and recognition setup consists of three state left-to-right HMMs with 32 Gaussian mixture components trained on the ESTER corpus [17]. The training corpus is comprised of 31 hours of broadcast news clean speech (no music overlaps and no telephone speech) from several French national radio programs. This corpus was manually transcribed. Feature vectors are extracted on half overlap 16 ms window. The vector consists of 12 MFCC, normalized energy, delta and delta delta (39 parameters). Context-independent acoustic models (39 monophones) were used. Initialization of models was done with automatic alignment of the Phase I training corpus using Baum-Welch re-estimation. This work was carried out with HTK [18].

A manual verification of the alignment has been performed in the initial portion allowed the calculation of the average spectrum of a set of consonants (/m/, /s/, /g/) and vowels (/e/, /i/,/~ɛ/). For normal speakers, this information is well differentiated in terms of frequency profile and energy, with extremes such as /m/ and /s/ (black and red in Figure 8). The /m/ consonant is characterized by a peak energy in low frequency, while the /s/ is dominated by high frequency contents.

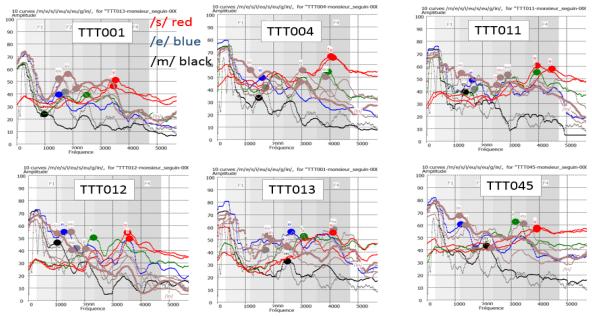

Figure 8: *Average energy profiles vs frequency (0-6 kHz) of vowels and consonants from normal speakers.*

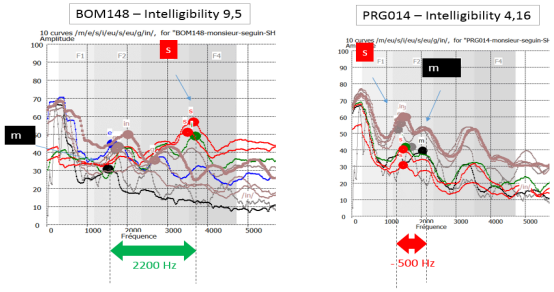

Figure 9: *Example of patients /s/m/ consonant distance in Hz.*

We determine the /s/m/ distance as shown in figure 10: the /s/ and /m/ maximum are detected within the F1..F4 formant region (that is adapted according to the patient gender). We evaluate the frequency distance between the 2 points in terms of frequency (X axis) and energy in dB (Y axis). For normal voices like Figure 9-left, both frequency and energy distance should be

positive between /s/ and /m/. However, in some case like shown in Figure 9-right, the distance may be negative. Table 1 gives the raw data information concerning the patient, gender, age, average F0 (extracted from the total read passage of 20 s), standard deviation of F0, normalized to 1.0, the intelligibility score, the /s/m/ distance in dB and Hz.

Note that the F0 standard deviation is not expressed in Hertz but as an indicator of pathology. An indicator higher that 1.0 means a standard deviation of F0 probability function below 3 notes, which indicates a monotone voice (OLD003, THS140). An indicator lower than 1.0 corresponds to a standard deviation higher than 3 notes, which indicates a normal voice (GAV008, ROA141).

A selection of patient's consonant and vowel profiles is given in Figure 10. Pathological voices are distinguished not only by the harmonic poverty already observed on the sustained /a/ but also an inability to differentiate the F3-F4 formants (NOF004, PRA142, PRG014) while 30% of the patients have spectrograms close to the normal voices (NIG145, OLD003).

The distance between the /s/ maximum (red spot in Figures 9 and 10) and /m/ maximum (black spot) within the F1-F4 formant bandwidth is extracted from the average spectrum, both in frequency (Hz) and energy (dB).

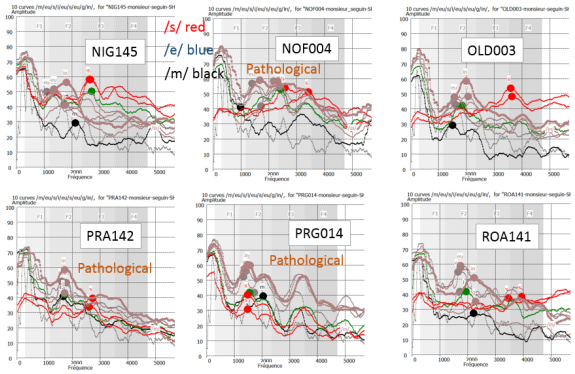

Figure 10: *Average energy profiles vs frequency (0-6 kHz) of vowels and consonants from patients.*

Table 1: List of patients with details of some basic voice features and /s/m/ consonant distance.

PATIENT	Sex	Age	Avg. F0	F0 STD Indic	INTEL Score 0..10	/s/M/ (dB)	/s/M/ (Hz)
ANC150	F	59	226	1,06	5	-4,24	366
BOM148	F	60	224	0,88	10	26,17	1981
BOT148	M	50	179	0,73	8	12,53	2562
CAA143	F	66	284	0,82	5	2,72	1960
COC005	F	65	215	0,69	7	13,28	2003
COC137	F	50	200	0,91	4	18,51	2347
FAS009	F	70	226	0,81	4	19,33	1335
FOD132	M	50	161	1,09	n.a	28,68	2670
GAV008	**F**	**61**	**192**	**0,47**	**5**	**28,91**	**2067**
GUY016	M	69	134	1,21	n.a	28,68	2670
JAR011	M	75	97	1,24	9	25,96	1981
JOP013	M	36	122	0,97	10	33,39	1809
LEM015	F	67	237	1,16	8	24,22	2239
LEY012	M	58	134	0,84	6	34,8	2369
NIG145	M	60	164	0,75	6	28,63	538
NIR008	M	61	146	1,15	4	28,63	538
NOF004	M	60	143	1,5	9	12,83	2433
OLD003	**M**	**63**	**139**	**1,64**	**n.a**	**24,76**	**2089**
PRA142	F	68	231	0,91	3	-1,18	1055
PRG014	M	66	101	1,38	4	0,82	-560
ROA141	**F**	**73**	**187**	**0,46**	**5**	**11,1**	**1723**
ROJ146	M	58	139	0,63	7	10,95	2476
RUM139	F	50	228	0,91	8	13,76	2218
SER007	M	74	137	1	1	-2,24	-258
SOM147	M	65	117	1,3	2	17,4	1443
THS140	**M**	**50**	**130**	**1,53**	**n.a**	**14,81**	**-43**

The correlation factor with intelligibility scores as presented earlier in Fig. 2 is around 0.67 for the distance in Hz, as illustrated in Fig. 11, but only 0.45 for the distance in dB.

We also attempted to correlate the indicator of harmonic poverty evaluated on a sustained /a/ and the intelligibility. Unfortunately, the correlation factor is very low (0.15), meaning that the spectral degradation of high-order harmonics do not forecast a lack of speech intelligibility.

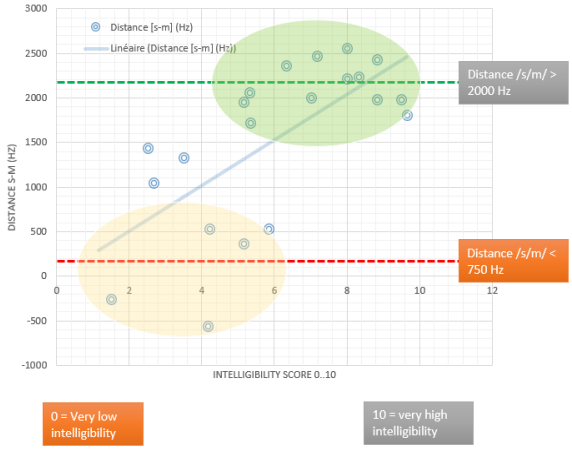

Figure 11: *Correlation between intelligibility score and /s/m/ consonant distance in Hz.*

4. Conclusion

This paper present several studies in order to reveal features that can be use in order to create a carcinologic speech severity index. This automatic C2SI measure will be well correlated with speech intelligibility score done by human listener.

The analysis of the voices of patients who were treated for oral cancer have revealed some tendencies: significant harmonic poverty associated with pitch and energy instability, a probability density function of the fundamental on a read passage very similar between normal read text, and low frequency differentiation between vowels and consonants for more than 50% of cases, particularly in zone F3-F4, leaving fears of confusion in the perception of pseudo-words and a loss of intelligibility.

This preliminary study has been conducted on the first third of the C2SI corpus. The experiments will be conducted on the other patient voices. The process of feature extraction has to be fully automatized. A C2SI index will be produced and will provide a speech intelligibility measure on the basis of the C2SI project protocol of recording. This index will provide an objective measure which may guide speech-language pathologists towards efficient voice therapy.

5. Acknowledgements

The authors would like to thank Dr. Woisard, coordinator of the C2SI project and all its members for sharing knowledge, methods, and patient's voice data base.

6. References

[1] Middag, C., Clapham, R., Van Son, R., & Martens, J. P. (2014). Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer. Computer speech & language, 28(2), 467-482.

[2] Ekteen E.C. et al. (2003). Comparison of voice characteristics following three different methods of treatment for laryngeal cancer. The Journal of otolaryngology, 32(4).

[3] Starmer, H. M., Tippett, D. C., & Webster, K. T. (2008). Effects of laryngeal cancer on voice and swallowing. Otolaryngologic Clinics of North America, 41(4), 793-818.

[4] Kraaijenga, S. A., et al. (2015). Prospective clinical study on long-term swallowing function and voice quality in advanced head and neck cancer patients treated with concurrent chemoradiotherapy and preventive swallowing exercises. European Archives of Oto-Rhino-Laryngology, 272(11), 3521-3531.

[5] Stone, D. et al. (2015). Voice Outcomes After Transoral Laser Microsurgery for Early Glottic Cancer—Considering Signal Type and Smoothed Cepstral Peak Prominence. Journal of Voice, 29(3), 370-381.

[6] C., Hilgers, F., et al. (2014). Developing automatic articulation, phonation and accent assessment techniques for speakers treated for advanced head and neck cancer. Speech Communication, 59, 44-54.

[7] Stemple, J. C., & Hapner, E. R. (2014). Voice therapy: clinical case studies. Plural Publishing

[8] https://www.irit.fr/recherches/SAMOVA/pagec2si.html

[9] Brierley, J. D. (2016), “TNM Classification of Malignant Tumours”, John Wiley & Sons

[10] Daudet, A. (1883) « Histoire de mes livres. Les Lettres de mon moulin », La Nouvelle Revue

[11] Menin-Sicard A., Sicard E., (2016) "Evaluation and Rehabilitation of the Voice - Clinical and Objective Approach", De Boeck Supérieur, ISBN 9782353273188, 288 pp.

[12] Gloinec Brendan, (2016). "Development of Methods to Characterize the Intelligibility of Patients Recovering from Throat Surgery", Master 2 report, INP-ENSEEIHT Toulouse, Sept. 2016

[13] Jacobson, B. H. et al. (1997) "The Voice Handicap Index (VHI): Development and Validation", American Journal of Speech-Language Pathology, Vol. 6, No. 3

[14] Fichaux-Bourin, P. et al. (2009) "Validation of a self-assessment for speech disorders (Phonation Handicap Index)" Revue de laryngologie, d'otologie et de rhinologie, vol. 130, no 1, pp. 45-51

[15] Leino, T. (2009), "Long-Term Average Spectrum in Screening of Voice Quality in Speech", Journal of Voice, Volume 23, Issue 6, pp. 671-676

[16] Rusz, J. et al. (2015). Speech disorders reflect differing pathophysiology in Parkinson’s disease, progressive supranuclear palsy and multiple system atrophy. Journal of neurology, 262(4), 992-1001.

[17] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier. 2005. The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In Proc. Interspeech. 1149– 1152.

[18] S. J. Young 1994. The HTK Hidden Markov Model Toolkit: Design and Philosophy. Entropic Cambridge Research Laboratory, Ltd 2 (1994), 2–44.