Language Resources and Evaluation manuscript No.
LREV-D-19-00107R1

C2SI corpus: a Database of Speech Disorder Productions to Assess Intelligibility and Quality of Life in Head and Neck Cancers

Virginie Woisard[1] · **Corine Astésano**[2] · **Mathieu Balaguer**[1] · **Jérôme Farinas**[5] · **Corinne Fredouille**[4] · **Pascal Gaillard**[2] · **Alain Ghio**[3] · **Laurence Giusti**[3] · **Imed Laaridh**[5] · **Muriel Lalain**[3] · **Benoît Lepage**[1] · **Julie Mauclair**[5] · **Olivier Nocaudie**[2] · **Julien Pinquier**[5] · **Gilles Pouchoulin**[3] · **Michèle Puech**[1] · **Danièle Robert**[3] · **Vincent Roger**[5]

Received: 2019/05/29 / Accepted: 2020/01/31

Abstract Within the framework of the Carcinologic Speech Severity Index (C2SI) INCa Project, we collected a large database of French speech recordings aiming at validating Disorder Severity Indexes. Such a database will be useful for measuring the impact of oral and pharyngeal cavity cancer on speech production. It will permit to assess patients Quality of Life after treatment. The database is composed of audio recordings from 134 sessions and associated metadata. Several intelligibility and comprehensibility levels of speech functions have been evaluated. Acoustics and prosody have been assessed. Perceptual evaluation rates from both naive and expert juries are being produced. Automatic analyzes are being carried out. It is intended to provide speech therapists and physicians with objective tools, which take into account the intelligibility and comprehensibility of patients which received cancer treatment (surgery and/or radiotherapy and/or chemotherapy). The aim of this paper is to justify the necessity of such a corpus and to present its data collection. This C2SI corpus will be available to the scientific community through the Scientific Interest Group Parolothèque.

[1] Toulouse University Hospital Centre
E-mail: woisard.v@chu-toulouse.fr

[2] Jean Jaures University, Toulouse, France
E-mail: astesano@univ-tlse2.fr

[3] Aix-Marseille Univ, CNRS UMR 7309, LPL, Aix-en-Provence, France
E-mail: alain.ghio@lpl-aix.fr

[4] Avignon University, LIA, Avignon, France
E-mail: corinne.fredouille@univ-avignon.fr

[5] Toulouse University, CNRS UMR 5505 IRIT, Toulouse, France
E-mail: jerome.farinas@irit.fr ***(corresponding author)***

Keywords speech intelligibility and comprehensibility · quality of life assessment · speech corpus · pathological speech

1 Introduction

The decreasing mortality in cancerology brings to light the necessity to reduce the impact of treatments on the Quality of Life (QoL) after cancer. That particularly concerns head and neck cancers (HNC), because their treatment can be mutilating and disabling. However, the usual tools for assessing QoL are not relevant for measuring the impact of the treatment on the main functions affected by the sequelae. And, there is a clear lack of uniform methods for assessing functional outcomes. Measuring the impact on one or several of the most altered functions after therapeutic care of a given tumoral localization would allow for: 1. completing the expression of the therapeutic outcomes by functional forecast index, 2. adjusting the treatment in order to reduce its functional consequences. For the HNC, it is mainly about impacts of (oral) communication and feeding (swallowing) [20]. QoL research has, to date, failed to provide health care professionals with clinically relevant and interpretable information that can guide treatment decisions. This has led researchers to attempt to make commonly used research tools more accessible to the clinicians. Health-Related Quality of Life (HRQoL) questionnaires reflect the disease impact or functional deficits on general well-being [6] by developing questions modules dedicated to the specific consequences of this disease or physiological function. But validated tools to measure the functional outcomes of carcinologic treatment are still missing, in particular for speech disorders. Some assessments are available for voice disorders in laryngeal cancer, but they are based on very poor tools for oral and pharyngeal cancers, dwelling on the articulation of speech rather than on the voice. Given that the usual tools to assess QoL are not relevant to measure the impact of the treatment on the main functions affected by the sequelae, and given that automatic speech processing tools are necessary for unbiased and objective assessments of communication deficiency caused by a speech disorder, we set out to develop a severity index of speech disorders describing the outcomes of therapeutic protocols supplementing the commonly used survival rates. The aim is to perform an audio recording of the patients speech and to compute the intelligibility of the utterances produced with the aim to get a score. Middag presented a new method that predicts running speech intelligibility in a robust way [19]. This method is text-independent and robust to differences in regional variations of Dutch/Flemish, hence robustly applicable to patients treated for HNC. Therefore, our hypothesis is that an automatic assessment technique can measure the impact of speech disorders on the communication abilities, by giving a severity index of speech for patients treated for HNC, more particularly for oral and pharyngeal cancers. We will name this index the Carcinologic Speech Severity Index (C2SI). Speech intelligibility 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, such as the familiarity effect experienced by clinicians with their patients' speech disorder, and the poor inter-judge reproducibility. Moreover, the scores do not accurately reflect listeners' comprehension. In order to develop and evaluate this C2SI, a project has been funded from 2014 to 2018 by the French National Cancer Institute (Grant INCa SHS 2014-135) including the following partners: (1) University Hospital Toulouse, (2) LPL laboratory from Aix-Marseille University, (3) Octogone-Lordat from Toulouse University, (4) LIA laboratory from Avignon University, (5) IRIT laboratory from Toulouse University. This C2SI project aims to create a speech corpus and to determine an automatic intelligibility measure. The C2SI corpus is presented

in this paper. The structure and the list of tasks performed by each speaker are presented in section 2. Section 3 presents the available material, and some statistics run on the corpus are reported in section 4.

2 Why did we build this corpus?

To cover the broad spectrum of intelligibility and comprehensibility aspects, we wanted to analyze speech distortions at different levels, involving several speaking tasks in order to apply complementary assessment methods. We also needed individual information through Quality of Life questionnaires.

2.1 Distortions during speech production

2.1.1 Voice signal

In general, low intelligibility is seen as a consequence of poor speech articulation, leading to the belief that there is a weak correlation between voice production and speech intelligibility. However, results from [25] indicate that "Patients with severe voice disorders showed very low intelligibility in spite of their intact articulation and prosody." A confirmation can be found in [24]. For instance, the capacity to hold a vowel more than five seconds in one breath is a minimal condition for a correct speech production. Recording such a sustained vowel (**AAA**) is a basic task linked to the aerodynamic/acoustic source performance of the speaker. This can also give indications on the speaker's breathing capacity. However, whereas measuring the voice level is really important in the case of laryngeal disease like, for instance, laryngeal cancers, this may not be the case with oral cavity cancers. If the relation "bad voice equal poor intelligibility" seems to be true, the reciprocal is false. Another way to state this is to say that a good voice is a necessary but not sufficient condition for good intelligibility.

2.1.2 Articulation quality

As a first proposal, we can give a definition of intelligibility of a speaker as *the performance by a listener to recognize the words and / or the sounds of the speech produced by the speaker*. We are close to the concept of articulation quality, and the idea is to take into account the accuracy of the phonetic realization. Sadly, intelligibility tests are performed with sentences or words extracted from a restricted list of items. The limitation of this type of test is the ability for listeners to restore the distorted sequences after some time of exposure to the same stimuli. This effect is emphasized when auditors have a strong knowledge of the words used in the test, and when these words are unambiguous and therefore strongly predictable, as in the FDA tests proposed by [8,9]. These restoration effects are clearly observable in speech-language pathologists who make such an extensive use of these lists that they eventually know them by heart. The bias associated with this knowledge, and therefore with the strong influence of the top-down perceptual mechanisms, results in an overvalued intelligibility score because the phonemic restoration of the listener makes production distortions opaque [31,27]. The solution we adopted consists in using large quantities of pseudo-words complying with the frequent phonotactic structures of the speakers native language, in order to completely neutralize the effects of lexicality, familiarization and learning of the items

by listeners [13]. Finally, listeners are confronted with a task that is similar to Acoustic-Phonetic Decoding (**DAP**) followed by a written transcription. The closer the transcription of the pseudo word to the target form, the better its intelligibility. We can make a quick calculation by simply counting the number of correctly recognized phonemes. We can also refine the method by counting the number of different phonetic features between the phonemes of the expected form and those of the transcribed form.

2.1.3 Continuous speech

In order to evaluate speech comprehension, it is important to go beyond the simple tests on isolated words.

1. We introduced a Sentence Verification Tasks (**SVT**) in order to assess the global comprehension of running speech. In this task, speakers read a set of sentences. The semantic content of each sentence can be true (ex: January is a winter month) or false (ex: January is a summer month). In the perception evaluation, participants are presented with a variety of utterances across several knowledge domains and have to decide as fast as possible if these statements are true or false [23]. The accuracy score and the response time are both used as indicators of the comprehension process. Indeed, when auditors need to understand the linguistic content of a message and perform an appropriate response [True or False], the quality of the acoustic-phonetic information of the speech signal plays an important role both in the speed and accuracy of the answer provided.
2. We also used a very common task, whereby speakers had to read a Short Text (**LEC**). This type of spoken communication is very useful because it integrates most of the linguistic levels (phonetic, lexical, syntactic, semantic) in a comparable way between speakers. It makes it possible to produce automatic phonetic alignments, even if the speech production is very altered. Speech rate, prosody, consonant and vowel precision, pauses and other speech features may be easily extracted and compared between the normal and patient groups.

2.1.4 Prosodic specific level

In spoken language, prosodic cues are at the interface of other linguistic levels and fulfill various functions in both message encoding and decoding. Prosody helps structuring utterances by indicating linguistic units' boundaries, thus fulfilling a syntactic function. It also serves the purpose of indicating sentence modality (assertion, interrogation, order...) by precise variations of the intonation contours. Prosodic devices such as focalization (strong accentual marking) are also used to highlight the central information of a message. Beyond these communicative functions, the coherent production of prosodic cues partakes in the fluency of utterances and their temporal organization. Prosody's multiple functions in speech, as well as its interaction with all levels of linguistic structuring, makes it an essential, indispensable feature of speech comprehensibility. Some models describe prosody as a tool for compensatory / palliative strategies to segmental alterations. In other words, speakers would tend to amplify the prosodic cues in their productions to make up for the loss of intelligibility / comprehensibility, may it be due to ambient noise or pronunciation difficulties (theory of speech adaptability, for communication optimization purposes [17]). The patients we focus on in this project have undergone treatment at the supra-laryngeal level of their anatomy (glossectomy, mandibulectomy for example). Theoretically, these treatments are not expected to impact on the production of the laryngeal flow or on prosodic indices, even

though some types of treatment may lead to a stiffening of the tissues beyond the treated area and, consequently, to a functional impairment of the larynx. We thus hypothesize that these speakers will obtain satisfactory results in the perceptual assessments with regard to the preservation of prosodic functions despite these peripheral risks: it will be difficult to distinguish between the patient and control groups solely on the basis of their scores, particularly during these tasks where the segmental information is negligible (syntax and modality). However, we believe that the articulatory damages on patients will have an impact on the response time of listeners' perception. Indeed, the alteration of patients' speech should result in increased difficulties of listeners' comprehension. The three prosodic tasks that we propose are designed to evaluate which structural functions of prosody are most affected by these types of cancer:

1. Modality Function (**MOD**): prosodic marking of assertion, question and injunction, by intonation contour shapes and directions.
2. Pragmatic Focus (**FOC**): this task required speakers to mark the pragmatic focus by highlighting the important information of an utterance by sole prosodic cues.
3. Syntactic Disambiguation (**SYN**): speakers had to solve syntactic ambiguity by prosodic means in syntagms composed of two nouns and an adjective, where the adjective either applied to both nouns (high syntactic attachment) or to the last noun (low syntactic attachment).

These tasks are taken from [2] who adapted [18] and [1] for clinical use. Speakers' capacity to properly use prosodic cues in these different tasks is then used for perceptual evaluation tests on naive, healthy listeners [21].

2.1.5 Spontaneous speech

In everyday speech, top-down effects are used to decode continuous speech. This is why spontaneous speech is very often used for assessing intelligibility [32]. In order to reduce speech predictability, we recorded patients and controls in a picture description task (**DES**), as well as in a free task where they had to spontaneously comment on a text they had read just before (**SPO**). Indeed, recording spontaneous speech can also be interesting to assess the comprehensibility of a text. But the evaluation of an index based on these recordings is not easy because semantics may widely vary. However, this task could also be analyzed in order to confirm the other indexes used in the perceptual analyzes.

2.2 Self Assessment questionnaires

Self-assessment questionnaires are used in practice to evaluate QoL in its several dimensions. The main generic quality of life questionnaire is the MOS-SF36 [30]. It is validated in all kinds of illnesses and explores physical as well as mental health disorders. Handicap self-assessment questionnaires were proposed for various functions of the upper aerodigestive tract (UADT). The Speech Handicap Index (SHI) for speech [26] is validated for HNC. The Phonation Handicap Index (PHI) is a similar tool for French, which is however validated for all kinds of speech production disorders [10]. The relationships between QoL questionnaires and Handicap questionnaires have often been analyzed, with the former being used to validate the contents of latter. Strong correlations (0.7 to 0.9) were computed between the SHI and the speech domain of the QoL questionnaire. This correlation is much lower, if not absent, regarding the other domains [4,7,29]. Because using the Handicap questionnaire

targeting a specific function is well correlated to the domains of the QoL questionnaires, we selected the generic QoL questionnaire (SF36) and the specific speech related handicap questionnaire (SHI and PHI) in order to integrate the communication dimension.

3 Speech Tasks

3.1 Sustained vowel **AAA**

These recordings consist in the production of 3 sustained /a/. A sustained vowel gives information about voice level, phonation time, stability, harmonics contents, noise, unvoiced segments, etc.

3.2 Pseudo-words **DAP**

Each speaker had to pronounce 52 pseudo-words. The pseudo-words have the following phonotactic structure : $C(C)_1V_1C(C)_2V_2$, where $C(C)_i$ was an isolated consonant or a consonant cluster. Such a combinatorial method made it possible to generate around 90000 pseudo-words. Each list contained the same amount of phonemes in C_1, V_1, C_2 and V_2 position, but in different combinations for each speaker [13]. Real words have been removed from the dictionary. See table 1 for a list example with these constraints.

Table 1 Example of a pseudo words list for DAP

spofo	stoumo	vurtant	muja	teilli	charou
chubra	blania	leba	quermant	neji	jarant
quindu	yainzou	yopta	froubin	finto	joucant
danfin	psitrin	squigu	tichu	ranto	pridi
sonin	gorquin	crazu	zouilla	vougou	grispi
trufru	plirbi	dinli	lanchin	banant	soublou
brussa	cliflu	glepa	zacrin	ruvo	pampou
floniant	drapro	manlo	nemou	nioucou	
poglu	bimpsi	guevant	finsi	nianscou	

To facilitate the production of pseudo-words by speakers, we used the software Perceval Lancelot[1]. The speaker was placed in front of a screen, and the pseudo-words were automatically displayed while a sound version was produced synchronously. This dual modality, visual and auditory, made it possible to limit reading errors. Given the large size of the corpus (89346 possible forms), the sound versions was computed using the Voxygen synthesis[2]. The recordings were then segmented semi-automatically, and each pseudo-word was automatically extracted in a separate audio file.

3.3 Sentences **SVT**

A list of 150 pairs of true/false sentences were created from [22], others from [33], while the remaining sentence pairs were created by the members of the C2SI project. These sen-

[1] `http://www.lpl-aix.fr/~lpldev/perceval/`

[2] `http://voxygen.fr/`

tences have a specific syntactic-semantic structure, whereby the true or false property can be checked only when the last lexical unit was produced (e.g. "Paris is the capital of France" *vs.* "Paris is the capital of Germany"). Consequently, it is necessary to decode and understand the whole sentence before coming up with the answer.

A set of 50 sentences selected from the list of 300 sentences was produced by each speaker.

3.4 Text **LEC**

The first paragraph of "La chèvre de M. Seguin", a tale by Alphonse Daudet, was read by the speakers. This text was chosen because it is long enough and it encompasses all French phonemes. It is also well known and widespread in clinical phonetics in France [14]. Here is the full plain text: "*Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon. Un beau matin, elles cassaient leur corde, s'en allaient dans la montagne, et là-haut le loup les mangeait. Ni les caresses de leur maitre, ni la peur du loup rien ne les retenait. C'était parait-il des chèvres indépendantes voulant à tout prix le grand air et la liberté.*"

3.5 Prosodic tasks

*3.5.1 Modality function **MOD***

The modality task consists in the production of ten identical sentences with 3 different modalities: assertion, question and injunction (*You eat pastas ?/./!*). Each speaker recorded 10 different scripts uttered with the 3 modalities. Each script was presented on a computer screen, with the expected prosodic modality indicated by either of the 3 punctuation marks ('.' '?' '!').

*3.5.2 Focus function **FOC***

In the focus task [2], speakers had to resolve a paradigmatic opposition (contrastive focus) between two words given in an auditorily presented sentence so as to prosodically highlight the relevant word ("*Did you see a duck or a pig in the garden?* with the written answer: *I saw a DUCK in the garden*). Each speaker recorded the same set of 20 sentences, for which they had to produce the proper focus as scripted, following the audio presentation of the question.

*3.5.3 Syntactic function **SYN***

The syntactic task [2, 1] consists of similar written scripts that only prosody can disambiguate. For example, in the sentence *les chevaux et les poneys blancs* (eg. *White horses and poneys*: note that the adjective in French is at the end of the sentence), the adjective *blancs* (eg. "*white*") can either apply to the second noun only (narrow scope) or to the two nouns (broad scope): prosodic cues such as final lengthening, pause and f0 excursions can give the proper syntactic parsing (either "*les chevaux // et les poneys blancs*" or "*les chevaux et les poneys // blancs*").

Each speaker recorded 13 scripts with two syntactic conditions (narrow *vs.* broad scope of adjective). The sentences were written on a computer screen, with the expected syntactic grouping indicated visually by vertical bars.

3.6 Picture description **DES**

The subject was asked to choose one amoung several pictures representing a similar scenery (the sea with boats). Each subject had to describe the picture to the examiner so that the latter could redraw it just on the basis of the oral explanations.

3.7 Spontaneous speech **SPO**

The patient had to give his/her opinion on the questionnaire that he/she has to fill out before the recording session. He/she had to speak for at least 3 minutes. This task allows us to collect spontaneous speech recordings with no constraint on the sentences.

4 Corpus description

4.1 Population

The number of patients to recruit in this database was estimated statistically on the following constraints : We expected the correlation between the automatic index and the perceived index given by the jury to be as high as 0.86 correlation, similar to the one achieved in the work done by the University of Ghent [19].

The size of the sample influences the precision of this estimation, a bigger sample bringing a bigger precision (characterized by a narrower confidence interval). To obtain a 95% confidence interval, the width of which is not superior to 0.15 around a coefficient of 0.8, it is necessary to recruit 94 patients. In September 2017, we recorded 134 sessions, that represents 87 patients and 40 control speakers. 7 patients were recorded twice. That is superior to the corpus used in [19], which contained recordings and perceptual evaluations of 55 patients with advanced Head and Neck Cancer who were treated with concomitant chemoradiotherapy. The patients were recruited in the three main departments of Toulouse managing patients with HNC (ENT department of the University Hospital, Cancerology department of the Institut Claudius Regaud (surgery and radiotherapy), Maxillofacial surgery department of the Toulouse University Hospital). They were selected from the lists of carcinologic follow-up consultations of these 3 departments. These departments are part of the University Institute of Cancer in Toulouse (IUC-T) and associated with the unit of "Oncoréhabilitation" which is located at the IUC-T Oncopole. These patients had to meet the following inclusion criteria:

- have a T1 to T4 cancer of the oral cavity and/or pharynx;
- have been treated by surgery and/or radiotherapy and/or chemotherapy;
- be more than 6 months after the end of treatment to ensure stability of the speech disorder, whether audible or not.

Similarly, the criteria for non-inclusion were to present another source of speech disorders (eg. stuttering) or to present cognitive or visual problems that are incompatible with the

assessment protocol design. These non-inclusion criteria were also used for the recruitment of the control population. Among the patients, 51 (59%) were men, and the mean age was 65.8 years old (range 36-87).

40 healthy controls (HC) were recruited. 18 control speakers (45%) were men. Figure 1 presents the age distribution of patients and controls. The control group's mean age was significantly different from the patients (56.9 years old, range 35-79, p=0.003 Mann-Whitney).

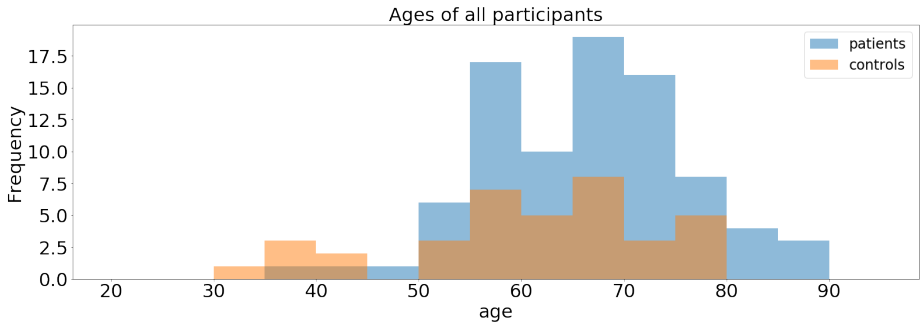

Fig. 1 Age distribution of the patients and controls groups

4.2 Metadata

Individual metadata were collected for each speaker. They comprise civility information such as age, gender, birth and area of residence (French department), as well as clinical information including the anatomical region affected by the cancerous lesion, values of T and N criteria from UICC Tumor/Node/Metastasis (TNM) classification [5] (the internationally accepted standard for cancer staging by the UICC journals), treatment type (surgery, radiotherapy, chemotherapy), time in months since the end of treatment.

The research protocol was reviewed by the Research Ethics Committee[3] (CER) from the University Hospital Centre of Toulouse. CER analyses ethical aspects of research protocols directly or indirectly involving humans. They approved the C2SI protocol on May 17th, 2016. A processing declaration which purpose is "the recording of the speech of patients treated for ENT cancer" was registered with the Commission Nationale de l'Informatique et des Libertés (CNIL) on July 24th, 2015 under number 1876994v0.

4.3 Questionnaires

The SHI and PHI health related quality of life questionnaires presented in 2.2 are given to the patients just before the audio recordings. The SHI is composed of 30 questions shared between two dimensions (symptoms and psycho-social). The PHI is a 15 items questionnaire with 3 dimensions (symptoms, functional consequences and emotional).

[3] `https://www.univ-toulouse.fr/actualites/comite-d-ethique-de-recherche-cer`

4.4 Recordings

The speakers were comfortably seated in an anechoic room in front of a computer (see figure 2). The computer was used to visually display instructions and corpus. For some tasks, the instructions were also produced with an auditory modality (ex: pseudo-words in DAP task). The recordings were made with a Neumann TLM 102 Cardioid Condenser Microphone connected to a FOSTEX digital recorder. The sampling rate was 48 kHz, which facilitates the downsampling to 16 kHz, usually used in automatic speech processing.

Fig. 2 Photography of the anechoic room for audio recordings

87 patients and 40 control speakers were recorded in the corpus. Unfortunately, despite a similar generic recording protocol, some patients and control speakers did not carry out all the tasks. Table 2 provides detailed information related to the tasks such as the number of speakers performing each task, the mean duration per recording as well as their total duration. It can be pointed out that 5 patients were recorded twice, performing all the protocol tasks. Similarly, 2 patients performed all the tasks except the spontaneous task, and a final patient was recorded twice on the AAA, LEC and DES tasks only. Regarding now the medical information, the inclusion criteria were balanced regarding tumor localization (see figure 3): 39% of oral cavity cancer (Floor of mouth, Tongue, Retromolar Area and Mandibula), and 61% of oropharyngeal cancer (Tonsil, Root of tongue, Soft Palate and when there is a larger extension OroPharynx).

Figure 4 presents the treatment distribution of patients. The most frequent treatment related to the size of the tumors is surgery (84%). The resection of the tumor (ChirT) is

Table 2 Information about speakers and tasks: number of speakers having carried out a given task (#FC: Female Control speakers - #FP: Female Patients - #MC: Male Control speakers - #MP: Male Patients), mean duration of recordings per task, total duration per task

Task	**Nb of speakers** #FC - #FP - #MC - #MP	**Mean duration per recording** (in seconds)	**Total duration** (in seconds)
Questionnaires			
SHI	6 - 30 - 5 - 50	-	-
PHI	6 - 29 - 5 - 48	-	-
SF36	12 - 35 - 8 - 49	-	-
Speech Tasks			
AAA	16 - 35 - 10 - 48	8.6	2978
DAP	21 - 36 - 18 - 44	188.6	24134
SVT	20 - 34 - 18 - 48	207.4	25920
LEC	21 - 33 - 18 - 47	33.1	3768
MOD	22 - 33 - 18 - 47	141.2	17645
FOC	22 - 33 - 18 - 46	192.7	25052
SYN	22 - 33 - 18 - 44	167.1	19889
DES	14 - 35 - 10 - 46	69.6	9397
SPO	15 - 34 - 10 - 43	66.0	7064

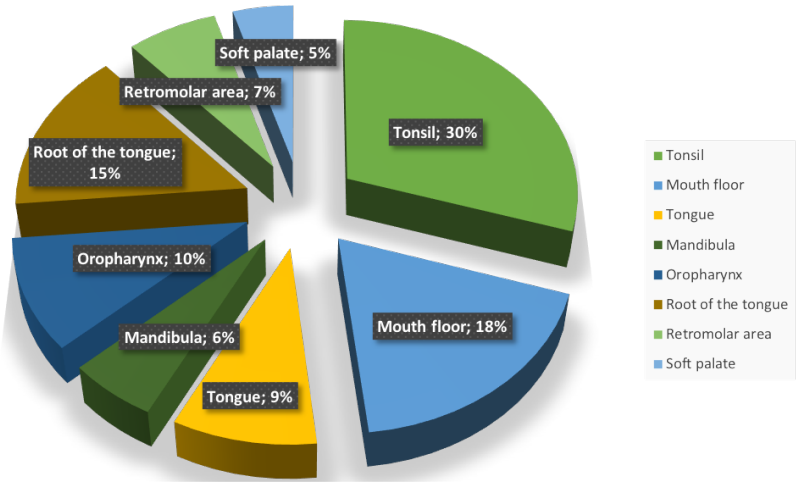

Fig. 3 Tumor Localization Distribution

associated with the node resection (ChirN) followed in 40% by a chemoradiotherapy (RT-chimio) and in 37% by radiotherapy (RT) only.

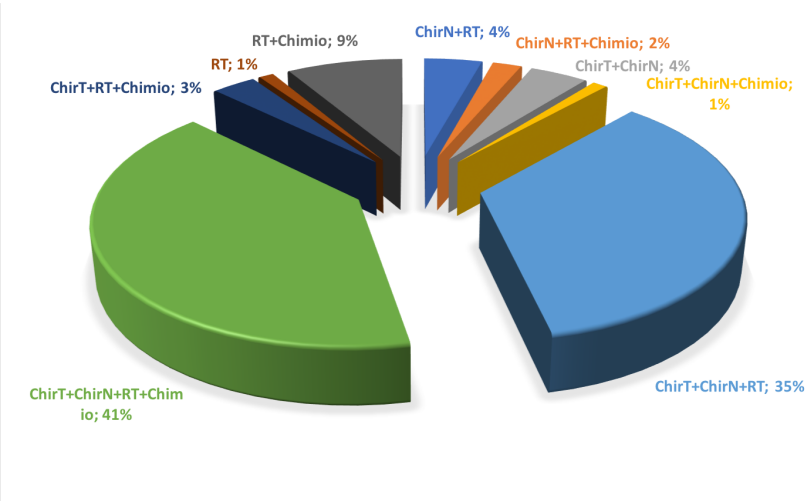

Fig. 4 Patients treatment distribution

5 Enrichment Data

5.1 Human perception evaluation

For the data DAP, MOD, SYN, FOC and SVT, the set of stimuli of all speakers was played back to a set of listeners via the Perceval Lancelot software[4]. We collected a large amount of perceptual data, allowing us to issue quantified indicators related to our data.

– DAP: All the 52 pseudo-words pronounced by every speaker of the database were transcribed 3 times. 40 naive listeners were involved in order to transcribe the 52 * 119 speakers = 6188 stimuli. Listeners were confronted with a task that can be considered as acoustic-phonetic decoding followed by a written transcription. The mean distance between the transcribed and expected response is considered as a score of (un)intelligibility. For the comparison operation, we used a Wagner-Fischer algorithm that integrates the phenomena of insertion, elision and substitution of units. In our case, this calculation of Levenshtein distance is not based on orthographic units but on phonemes [13]. Indeed, on the orthographic forms, in a traditional way, the distance between 2 graphemes is null if they are equal and is equal to 1 if they are different. In the case of phonemes, it is possible to introduce more subtle nuances because, for example, we can consider that a confusion between two vowels does not have the same weight as between a vowel and a voiceless consonant. We used the phonetic features theory to establish the local distance. In our preliminary results, we validated this measure which is discriminant between healthy speakers *vs.* patients. The measure called Perceived Phonological Deviation (PPD) is strongly correlated with the clinical severity index. Moreover, experiments show that the method is not biased by a learning effect by the listeners.

[4] `http://www.lpl-aix.fr/~lpldev/perceval/`

In order to perceptually evaluate 3 times each recorded sentence for SVT (4816 sentences), FOC (1969 sentences, MOD (2860 sentences) and SYN (2513 sentences) tasks, the recordings were presented to 147 naive listeners:

- MOD: the recordings were presented to naive listeners, who had to recognize which modality was intended, between assertion, question and injunction [21].

- FOC: Each sentence previously recorded was thereafter associated with a congruous ("*Qu'as-tu vu dans le jardin, un cochon ou un canard ?*" / eg. "*What did you see in the garden, a pig or a duck?*") or incongruous ("*Où as-tu vu un canard, dans le jardin ou dans la cour ?*" eg. "*Where did you see a duck in the garden or in the yard?*") question. Listeners had to judge whether the perceived focus was congruous or incongruous in the manipulated dialogues.

- SYN: Each recorded sentence was presented to naive listeners who had to choose between two pictures representing either one or the other syntactic reading (narrow *vs.* broad scope of adjective).

- SVT: The sentences were evaluated by 3 naive listeners who had to judge whether the sentence presents a true fact or an incorrect one. This produces an indicator based of the global comprehensibility of the sentence recorded.

A perception score was thereafter calculated for each speaker*task (SYN, FOC, MOD and SVT). These scores ranging from 0 to 3 correspond to the mean of each perceptual evaluation obtained during the test. Mean listeners reaction times were also calculated. These perceptual scores were correlated to an index of severity (estimated by 5 healthcare professionals) as reported in [21]. It appears that the mean SVT score is strongly correlated to this index of severity (r = .81, p.<.001) and thus can serve as a rating of the global comprehensibility of speakers whereas prosodic tasks mean scores are moderately correlated to the severity index (FOC : r = .56, p<.001 , MOD : r = .44, p<.05 ; SYN : r = .53, p<.001), indicating that tested prosodic functions seem overall preserved in the patients speech.

- LEC and DES: With the data obtained on read and spontaneous speech, an index of severity (alteration of speech signal) and subjective intelligibility was produced by a set of six experts on a scale from 0 (the strongest alteration) to 10 (perfect speech). In order to assess for inter-judge reliability, an Interclass Correlation Coefficient (ICC) was calculated. The degree of concordance between the jury ratings is therefore good ($r > 0.69$) for the set of tasks. The jury constituted by the six speech-language expert jurors is therefore homogeneous.

Although these different tasks give highly correlated results ($r > 0.8$), the intent to evaluate speech severity (alteration of the speech signal) favors a score distribution offering a better metric. The severity is on average more impacted by an oral location (5.44, sd 2.47) than oropharyngeal (6.46, sd 2.24). The semi-spontaneous speech tends to reduce the ceiling effect of the severity measure compared to the speech read (average scores at 6.06 over 10 in image description, and 6.51 over 10 in reading).

Perceptive measurement of the severity of speech disorder on semi-spontaneous speech seems to be the clinically most relevant score in the evaluation of speech disorders after cancer treatment of the oral cavity or oropharynx. More details on this perceptual task could be found in [3]. Figure 5 presents the distribution of intelligibility and severity scores across subjects on DES task. The figure is sorted by increased scores of intelligibility.

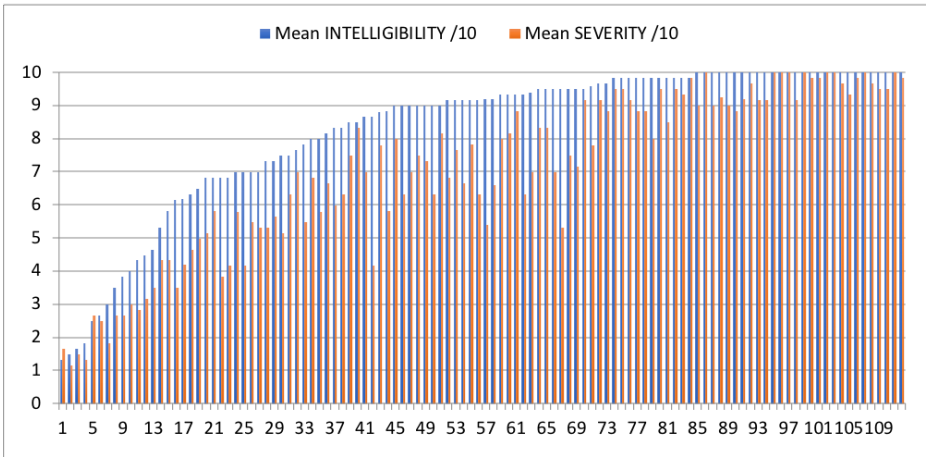

Fig. 5 Distribution of intelligibility and severity scores on DES task in ordinate and subjects (by increased scores of intelligibility) in abscissa

5.2 Automatic processing

In order to enrich the C2SI corpus, notably for phonetic analysis related to speech disorder analysis, audio recordings related to the LEC task was segmented automatically at the phone level thanks to a forced-alignment system. The latter takes as inputs the sequence of words pronounced in a speech utterance, and a phonetized lexicon of words coupled with different phonological variants, based on a set of 37 French phones. The forced alignment is based on a Viterbi decoding and graph-search algorithms, the core of which is the acoustic modeling of each phone, based on Hidden Markov Models (HMM). A 3-state context-independent HMM topology is used to model each phone. The HMM-based models are built thanks to the Maximum Likelihood Estimate paradigm from about 200 hours of French radiophonic speech recordings [12]. These models are speaker independent, however a three-iteration MAP adaptation is applied to all the HMM parameters to get speaker-dependent models. Acoustic vectors consist of 12 Perceptual Linear Prediction coefficients plus the energy, plus their delta and delta-delta coefficients.
It is important to note that the input sequences of words come from the original text that speakers had to read for the LEC task. Indeed, no manual orthographic transcription per recording had been performed by human listeners. Therefore, alignment errors could be possible due to word repetitions, omissions, substitutions, or deletions. Nevertheless, depending on the targeted phonetic analysis, a manual phone segmentation correction could be envisaged from the automatic outputs, which may bring a large gain of time.
Thus, this corpus enrichment results in one pair of start and end boundaries per phone present in all the speech recordings associated with the LEC task, packaged into TextGrid format files.

6 Conclusions and future work

In this paper, we have presented the design and recording of a corpus of 127 speakers, which allows us to consider the automatic production of indexes with a high level of correlation.

During the constitution of the corpus, we faced several issues. Considering DAP task, patients recordings were initially achieved, using only a visual presentation of the DAP items and the pseudo-word was simultaneously read aloud by the experimenter. However, because the phonological construction of the items sometimes allows different possible pronunciations, this configuration could have modified the speakers repetition. To cope with this drawback, we replaced the aloud reading of the experimenter with a recorded synthesized voice for each item to standardize its pronunciation and to limit the potential biases. Furthermore, some tasks were considered as particularly hard to understand and to achieve by the patients (SYN, for example): the impact of these perceived difficulties will have to be checked and studied during the analysis of the results. Perceptual evaluations are in progress in order to complete the usable metadata, and to obtain reliable intelligibility/comprehensibility scores, which will be compared to self-assessed quality of life scores. We are also working now on extracting information from the different recordings in order to analyze them and to produce automatic indexes [28, 16, 15, 11].

Our main goal is to get objective judgments, which can help speech therapists and physicians in clinical practice. Data will be available to the scientific community through the GIS Parolotheque[5]: a scientific structure ("Groupement d'Intret Scientifique") which purpose is to facilitate access and research on pathological speech recordings (like the tumor library thomorothèque for access to cancer cell samples). The data come from hospital structures in a pseudo-anonymized form (waveforms cannot be totally anonymized by definition, but all metadata from the hospital are cleaned). The GIS is responsible for the storage, legal aspect and allocation of access to scientists in the context of research projects. Access to data and metadata is therefore facilitated and accelerated compared to the traditional approaches that are currently required to participate in this kind of research. The GIS is currently in a signing phase and operational implementation is expected to start in 2020.

Acknowledgements Grant 2014-135 from Institut National pour le CAncer (INCa) in 2014, Sciences Humaines et Sociales, épidémiologie et Santé Publique call. Lead by Pr Virginie Woisard at University Hospital of Toulouse and Grant ANR-18-CE45-0008 from The French National Research Agency in 2018 RUGBI project "Improving the measurement of intelligibility of pathological production disorders impaired speech" lead by Jérôme Farinas at IRIT. We thank the company Voxygen[6] for providing us with their speech synthesis platform necessary for the realization of the corpus DAP.

References

1. Astésano, C., Bard, E.G., Turk, A.: Structural influences on initial accent placement in french. Language and Speech **50**(3), 423–446 (2007)
2. Aura, K.: Protocole d'évaluation du langage fondé sur le traitement de fonctions prosodiques : étude exploratoire de deux patients atteints de gliomes de bas grade en contexte péri-opératoire. Ph.D. thesis, Université Toulouse 2 (2012). URL `http://www.theses.fr/2012TOU20110/document`
3. Balaguer, M., Boisguerin, A., Galtier, A., Gaillard, N., Puech, M., Woisard, V.: Assessment of impairment of intelligibility and of speech signal after oral cavity and oropharynx cancer. Annales fran caises d'oto-rhino-laryngologie et de pathologie cervico-faciale **136**(5), 355–359 (2019). URL `https://doi.org/10.1016/j.anorl.2019.05.012`
4. Borggreven, P.A., Aaronson, N.K., Verdonck-de Leeuw, I.M., Muller, M.J., Heiligers, M.L., de Bree, R., Langendijk, J.A., Leemans, C.R.: Quality of life after surgical treatment for oral and oropharyngeal cancer: a prospective longitudinal assessment of patients reconstructed by a microvascular flap. Oral oncology **43**(10), 1034–1042 (2007)

[5] `https://www.parolotheque.fr`

[6] `http://voxygen.fr/`

5. Brierley, J.D., Gospodarowicz, M.K., Wittekind, C.: TNM classification of malignant tumours. John Wiley & Sons (2016)
6. Cardol, M., Brandsma, J., De Groot, I., van den BOSOE, G., De Haan, R., De Jong, B.: Handicap questionnaires: what do they assess? Disability and rehabilitation **21**(3), 97–105 (1999)
7. Dwivedi, R.C., St. Rose, S., Roe, J.W., Chisholm, E., Elmiyeh, B., Nutting, C.M., Clarke, P.M., Kerawala, C.J., Rhys-Evans, P.H., Harrington, K.J., et al.: First report on the reliability and validity of speech handicap index in native english-speaking patients with head and neck cancer. Head & neck **33**(3), 341–348 (2011)
8. Enderby, P.M.: Frenchay dysarthria assessment. Pro-ed (1983)
9. Enderby, P.M., Palmer, R.: FDA-2: Frenchay Dysarthria Assessment: Examiner's Manual. Pro-ed (2008)
10. Fichaux-Bourin, P., Woisard, V., Grand, S., Puech, M., Bodin, S.: Validation of a self assessment for speech disorders (phonation handicap index). Revue de laryngologie-otologie-rhinologie **130**(1), 45–51 (2009)
11. Fredouille, C., Ghio, A., Laaridh, I., Lalain, M., Woisard, V.: Acoustic-phonetic decoding for speech intelligibility evaluation in the context of head and neck cancers. In: Proceedings of Intl Congress of Phonetic Sciences (ICPhS'19). Melbourne, Australia (2019)
12. Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.F., Gravier, G.: The ester phase ii evaluation campaign for the rich transcription of french broadcast news. In: Ninth European Conference on Speech Communication and Technology (2005)
13. Ghio, A., Lalain, M., Giusti, L., Pouchoulin, G., Robert, D., Rebourg, M., Fredouille, C., Laaridh, I., Woisard, V.: Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique. In: XXXII éme Journées d'Etudes sur la Parole (2018). DOI 10.21437/JEP.2018-33. URL `https://hal.archives-ouvertes.fr/hal-01770161/file/190996.pdf`
14. Ghio, A., Pouchoulin, G., Teston, B., Pinto, S., Fredouille, C., De Looze, C., Robert, D., Viallet, F., Giovanni, A.: How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers? Speech Communication **54**(5), 664–679 (2012)
15. Laaridh, I., Fredouille, C., Ghio, A., Lalain, M., Woisard, V.: Automatic Evaluation of Speech Intelligibility Based on i-vectors in the Context of Head and Neck Cancers. In: Interspeech, pp. 2943–2947. ISCA, Hyderabad, India (2018). DOI 10.21437/interspeech.2018-1266. URL `https://hal.archives-ouvertes.fr/hal-01962170`
16. Laaridh, I., Kheder, W.B., Fredouille, C., Meunier, C.: Automatic prediction of speech evaluation metrics for dysarthric speech. In: Proc. Interspeech, pp. 1834–1838 (2017)
17. Lindblom, B.: Explaining phonetic variation: A sketch of the h&h theory. In: Speech production and speech modelling, vol. 55, pp. 403–439. Springer, Dordrecht (1990). URL `https://doi.org/10.1007/978-94-009-2037-8_16`
18. Magne, C., Astésano, C., Lacheret-Dujour, A., Morel, M., Alter, K., Besson, M.: On-line processing of pop-out words in spoken french dialogues. Journal of cognitive neuroscience **17**(5), 740–756 (2005)
19. Middag, C., Clapham, R., Van Son, R., Martens, J.P.: Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer. Computer speech & language **28**(2), 467–482 (2014)
20. MlynareK, A.M., Rieger, J.M., Harris, J.R., O'Connell, D.A., Al-Qahtani, K.H., Ansari, K., Chau, J., Seikaly, H.: Methods of Functional Outcomes Assessment following Treatment of Oral and Oropharyngeal Cancer: Review of the Literature. Journal of otolaryngology - head and neck surgery **37**(1), 2–10 (2008)
21. Nocaudie, O., Astésano, C., Ghio, A., Lalain, M., Woisard, V.: Evaluation de la compréhensibilité et conservation des fonctions prosodiques en perception de la parole de patients post traitement de cancers de la cavité buccale et du pharynx. In: XXXIIe Journées d'Etudes sur la Parole, pp. 196–204 (2018)
22. Pisoni, D.B., Dedina, M.J.: Comprehension of digitally encoded natural speech using a sentence verification task: a first report. Tech. Rep. Progress report 12, Indiana University (1986)
23. Pisoni, D.B., Manous, L.M., Dedina, M.J.: Comprehension of natural and synthetic speech: Effects of predictability on the verification of sentences controlled for intelligibility. Computer speech & language **2**(3-4), 303–320 (1987)
24. Porcaro, C., Evitts, P., King, N., Hood, C., Campbell, E., White, L., Veraguas, J.: Effect of dysphonia and cognitive-perceptual listener strategies on speech intelligibility. Journal of Voice **in press** (2019). DOI https://doi.org/10.1016/j.jvoice.2019.03.013
25. Pyo Hwa Young, S.H.S.: A study of speech intelligibility affected by voice quality degradation. Commun Sci Disord **12**(2), 256–278 (2007). URL `http://www.e-csd.org/journal/view.php?number=326`
26. Rinkel, R.N., Leeuw, I.M.V.d., van Reij, E.J., Aaronson, N.K., Leemans, C.R.: Speech handicap index in patients with oral and pharyngeal cancer: better understanding of patients' complaints. Head & Neck: Journal for the Sciences and Specialties of the Head and Neck **30**(7), 868–874 (2008)

27. Samuel, A.G.: Phonemic restoration: insights from a new methodology. Journal of Experimental Psychology: General **110**(4), 474 (1981)
28. Sicard, E., Mauclair, J., Woisard, V.: Etude de paramètres acoustiques des voix de patients traités pour un cancer orl dans le cadre du projet c2si. In: 7èmes Journées de Phonétique Clinique (2017)
29. Thomas, L., Jones, T.M., Tandon, S., Carding, P., Lowe, D., Rogers, S.: Speech and voice outcomes in oropharyngeal cancer and evaluation of the university of washington quality of life speech domain. Clinical Otolaryngology **34**(1), 34–42 (2009)
30. Ware Jr, J.E., Sherbourne, C.D.: The mos 36-item short-form health survey (sf-36): I. conceptual framework and item selection. Medical care pp. 473–483 (1992)
31. Warren, R.M., Warren, R.P.: Auditory illusions and confusions. Scientific American **223**(6), 30–37 (1970)
32. Woisard, V., Espesser, R., Ghio, A., Duez, D.: De l'intelligibilité à la compréhensibilité de la parole, quelles mesures en pratique clinique? Revue de Laryngologie Otologie Rhinologie **1**(134), 27–33 (2013)
33. Zumbiehl, O.: Evaluation perceptive des dysphonies par la sentence verification task. Master's thesis, Université Aix-Marseille (2010). Mémoire dOrthophonie (dir. : Cavé, C. and Ghio, Alain)