Validation of an intelligibility test based on acoustic-phonetic decoding of pseudo-words: overall results from patients with cancer of the oral cavity and the oropharynx

Alain Ghio, Muriel Lalain, Marie Rebourg, Anna Marczyk, Corinne Fredouille, Virginie Woisard

Folia Phoniatrica et Logopaedica

Research Article

Folia Phoniatr Logop
DOI: 10.1159/000519427

Received: April 26, 2021
Accepted: August 26, 2021
Published online: September 21, 2021

Validation of an Intelligibility Test Based on Acoustic-Phonetic Decoding of Pseudo-Words: Overall Results from Patients with Cancer of the Oral Cavity and the Oropharynx

Alain Ghio[a] Muriel Lalain[a] Marie Rebourg[a] Anna Marczyk[a]
Corinne Fredouille[b] Virginie Woisard[c]

[a]Aix-Marseille University, CNRS, LPL, UMR 7309, Aix-en-Provence, France; [b]LIA, Avignon University, Avignon, France; [c]Service ORL, Centre Hospitalo-Universitaire Toulouse Hôpital Larrey, URI Octogone-Lordat, Université Toulouse II, Toulouse, France

Keywords
Intelligibility · Head and neck cancer · Speech production deficit · Speech disorder · Speech perception

Abstract
Objectives: Loss of intelligibility is a major complaint for patients with speech disorders, as it affects their everyday communication and thus contributes to a decrease in their quality of life. Several tests are available to measure intelligibility, but these tests do not take into account the evaluators' ability to restore distorted sequences. Due to this ability, the evaluator will tend to recognize words despite phonetic distortions, and speech production deficit can go undetected. The results of these tests therefore overestimate the intelligibility of patients and may mask real functional limitations. We propose a new test which uses a large number of pseudowords in order to neutralize the unwanted perceptual effects that cause this overestimation. The purpose of this test is to measure the speech production deficit. It is not intended to assess the communication deficit. Our objective is to validate this test based on acoustic-phonetic decoding of productions from patients with speech disorders. ***Materials and Methods:*** We tested this method with a population of 39 healthy participants and 78 post-treatment patients with cancers of the oral cavity and the oropharynx (HNC patients). Each speaker produced 52 pseudowords taken from randomly generated lists from large common dictionary, each list of 52 pseudowords containing the same number of phonemes. Forty everyday listeners then transcribed these productions. The orthographic transcriptions were phonetized and compared to the expected phonetic forms. An algorithm provided a Perceived Phonological Deviation score (PPD) based on the number of features that differed between the expected forms and the transcribed items. The PPD thus provided a score representing the loss of intelligibility. ***Results:*** The 39 participants in the control group demonstrated significantly lower PPD scores compared to the 41 patients with a T1T2 tumor size or compared to the 37 patients with a T3T4 tumor size. The differences between the three groups were significant. If we use the PPD as a predictor to identify patients versus control group subjects, the AUC of the ROC curve is equal to 0.94, which corresponds to an outstanding group separability. A PPD threshold at 0.6 features per phoneme is the boundary between normal and dysfunctional speech. The analysis showed a close correlation between the PPD and a clinical judgment of the disorder severity obtained from experts. ***Conclusion:*** This test ap-

karger@karger.com
www.karger.com/fpl

Correspondence to:
Alain Ghio, alain.ghio@lpl-aix.fr

pears to be effective in measuring the intelligibility of speakers at a phonological level, in particular in the case of head and neck cancers.

Introduction

Rationale

Regardless of their etiology, speech production disorders can cause a significant communication deficit that has a major impact on everyday life. In a study on a population with head and neck cancer (HNC), Meyer et al. [1] found a significant correlation between speech intelligibility and some aspects of quality of life measured by self-assessment. The authors noted an association between intelligibility and quality of life, remarking that "this disease may disrupt daily activities as a result of altered speech" [2]. A functional communication deficit, often the chief complaint of surviving patients, is usually examined within a 2-fold clinical assessment. The impairment, a "loss or abnormality of anatomical structure" in the case of cancer according to the World Health Organization (WHO) [3], is estimated by examining components of speech production including respiration, phonation, velopharyngeal function, and oral articulatory structures (jaw, tongue, and lips) [4]. Secondly, the evaluation aims to precisely identify functional limitations, or the "the lack of ability to perform an action in the manner considered normal that results from impairment" [3]. At this level of assessment, the goal is to measure how well HNC patients can use their preserved articulators to produce the intended acoustic output. As proposed by Yorkston et al. [5], "measures such as speech intelligibility are targets of assessment" for this level, indicating functional limitations.

Speech intelligibility assessment relies on multiple tools. The correlation analysis between several metrics can provide a comprehensive description of impairment profiles. The pseudoword-based material we present here is not designed to be used *instead* of word-based or utterance-based materials, rather, it is complementary to other types of assessment tools that should be used according to its potential and the advantages it presents over other types of elicitation materials. It is thus particularly adapted whenever there is a need to minimize the listener's effects on intelligibility measurements. It is also recommendable if lexical effects must be neutralized to isolate post-lexical sources of intelligibility reduction. The test that we propose specifically targets the perceptual impact of the "loss or abnormality of anatomical structure" [3], which is a first element of the ENT assessment relating to HNC. Of course, it can be supplemented by more targeted tests on the communication handicap. Finally, phonetically balanced pseudowords are advantageous if a systematic phonetic analysis of mistranscriptions is proposed to explore the source of intelligibility reduction, to relate error patterns to specific etiologies, or guide therapy intervention.

In the present work, we distinguish the concept of intelligibility from comprehensibility. We adopt a strict definition of intelligibility as "the amount of speech understood only from signal-dependent information," which is a concept proposed by Lindblom [6] in his model of spoken communication. Currently available tests of intelligibility contain certain drawbacks, which led us to develop a new tool based on acoustic-phonetic decoding to measure functional limitations, in particular speech production deficit rather than communication disorder. Our specific goal is to validate our protocol in order to make it available for speech disorder assessments in clinical settings. Our intelligibility tool was tested on patients with speech conditions resulting from cancer of the oral cavity and the oropharynx, but it could be generalizable to patients with any type of articulatory-motor speech disorder.

Speech Intelligibility versus Speech Comprehensibility

According to Yorkston et al. [5], compared to intelligibility, comprehensibility includes additional factors such as semantic context, syntactic context, situational cues, orthographic supplementation, gestures, or pictures. In the WHO classification [3], comprehensibility relates to a level of disability, which is a limitation in performing social activities within a social and physical environment. Comprehensibility indicates the "adequacy of speech performance in a social context" [5]; it refers to the concept of communication efficiency in everyday life.

The Lindblom model [6] of spoken communication describes two sources of information which are necessary for comprehension. The first is "signal-dependent information," which is extracted from the speech signal by a bottom-up process. This process, called "acoustic-phonetic decoding," consists of identifying phonemes from the speech signal. Since phonemes are the smallest units used to express meaning, phonemes can be considered as the basic units of speech intelligibility. Acoustic-phonetic decoding is therefore the fundamental process used in perceptual measurements of speech intelligibility.

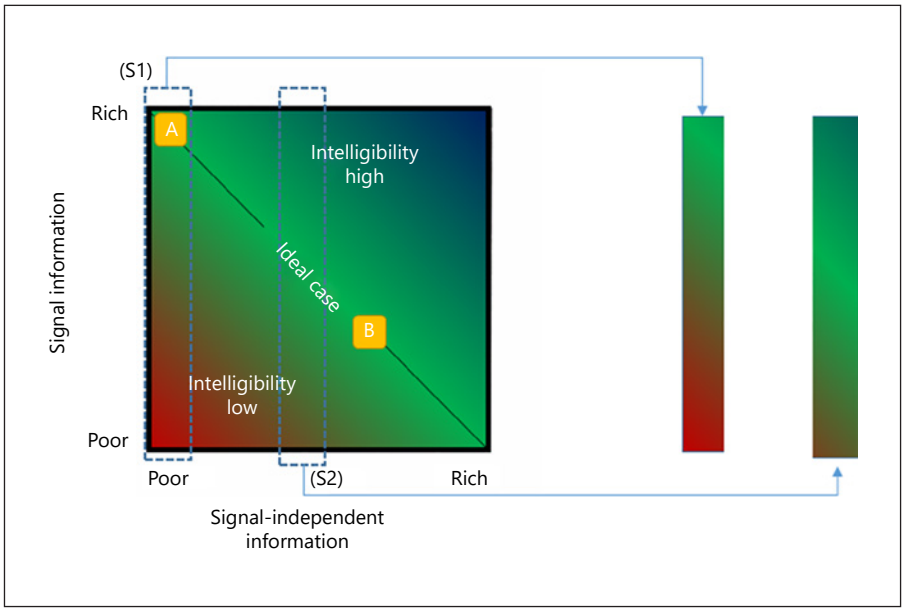


Fig. 1. The Lindblom model of spoken communication. The intelligibility level varies from low (red) to high (blue) with an intermediate ideal case (green). Landmark A: message without signal-independent information. Landmark B: message with important signal-independent information.

The second source of information is "signal-independent." This source is the result of top-down processes where the listener constructs the message from all the information at their disposal at different levels: knowledge of the lexicon in general, knowledge of the communicative context that restrains the activation of lexical units, shared knowledge between speakers, encyclopedic knowledge, the psychosocial context, etc. This cognitive process is strongly linked to comprehensibility, which is defined by Fontan et al. [7] as "the integration of both acoustic-phonetic information and all relevant information independent of the signal in order to understand a spoken message in a particular communicative situation." If we consider comprehensibility to be based on both signal-dependent and signal-independent sources of information, we can define intelligibility as the amount of speech understood only from signal-dependent information.

The Communication Model of Lindblom

In his model of spoken communication (Fig. 1), Lindblom [6] argues that when signal-dependent information is precise, the listener is able to understand the message without signal-independent information (Fig. 1, point A). On the other hand, when the signal-dependent information is insufficient, signal-independent information becomes crucial to understanding the speaker's message. If a speaker with a speech production disorder provides listeners with an imprecise signal, the speaker will try to compensate, with help from the interlocutor, by increasing the amount of signal-independent information to fill the gaps left by incomplete or compromised signal-dependent information (Fig. 1, point B). In everyday life, these processes are essential and widely used in communicative situations involving patients. Phonological distortions due to articulation/phonatory imprecision are generally compensated for in the context of natural communication. However, clinical assessments of speech disorders need to focus on measuring speaker performance while minimizing listener and context-related effects because these variations, which are external to the speaker, can be considered as measurement noise.

Current Intelligibility Tests

In a clinical context, there are currently two ways to rate speech intelligibility [8]: in the first, experts subjectively estimate intelligibility level in a scaling task, and in the second, intelligibility is assessed based on the identification of sentences or words from reference lists. We focus on the second method, which is generally described as more objective [9].

A number of intelligibility tests based on item identification have been developed. Most of them have been designed for dysarthria assessment, but they can also be used for other speech production disorders. A few examples include the "Single Word Intelligibility Test" by Tikofsky [10], the "Frenchay Dysarthria Assessment" by Enderby [4], the "Assessment of Intelligibility in Dysarthric Speakers" by Yorkston and Beukelman [11], the "Multiple Word Intelligibility Test" by Kent et al. [8], and the "Sentence Intelligibility Test" by Yorkston et al. [12].

In Europe, Enderby's FDA test is probably the most widely used in its first version [4] or second version [13]. It has been adapted to many languages, including French, German, Dutch, Norwegian, Swedish, Finnish, Catalan, Castilian, Portuguese, and Italian. The FDA test is particularly interesting because it evaluates the patient both in terms of functional limitations through intelligibility testing and in terms of impairment with an analytic grid that includes items on reflexes, breathing, larynx, lips, palate, and tongue. The intelligibility part of the test is composed of a list of 50 words and 50 sentences in the first version [4] and 116 words and 50 sentences in the second version [13]. During the assessment, the patient reads ten words and ten sentences aloud, and the examiner writes down what they hear. Then the examiner counts the number of words and sentences which were correctly recognized. Combined with a subjective analysis of speech (based on 5 min of conversation), this test provides an intelligibility score.

Current Limitations in Intelligibility Assessment

The main limitation of this type of evaluation lies in the fact that listeners involved in perceptual assessments are able to restore distorted sequences. For example, if a patient pronounces the word "topic" ['tɒpɪk] as *thovig* [θɒvɪg] and the listener recognizes the word "topic," the answer is noted as correct, and no problem is detected. However, three phonemes out of five were produced with major distortions. This restoration effect has been demonstrated by several studies, such as Warren and Warren [14], Ganong [15], and Samuel [16]. The effect is even stronger if the listener is familiar with the words used in the test and if these words are unambiguous and therefore highly predictable. This is generally the case for speech-language pathologists who use these lists extensively and end up knowing them by heart. For instance, the SI-BECD list in French [17] has only fifty words in its first version and a hundred in its second version [18]. Bias linked to familiarity with words [19] results in an overestimated intelligibility score because the listener's phonemic restoration masks distortions in patients' productions. The listener only notices strong alterations, and thus the test has low sensitivity. It is clear that top-down listener-related information influences the perceptual outcome in current evaluation batteries. In other words, current evaluations assess patients' comprehensibility rather than their intelligibility as defined above, even if the assessment is based on isolated words or short sentences.

In the Lindblom model described above, previous knowledge or familiarity with words can be translated as significant signal-independent information. In Figure 1, this situation is plotted as S2. If we isolate the S2 condition (right part of the Figure) and make a visual assessment, we can see that the largest part of the square is uniform, which means that in this situation, the intelligibility level (the color) is uniform and poorly correlated with the signal-dependent information. In other words, the result has little to do with the speaker, which is inconvenient for patient assessments. The red color, which indicates only a significant functional limitation, appears if the level of signal information is very low.

Our Proposal

Returning to the Lindblom model in Figure 1, we propose to move the situation to the S1 position, where signal-independent information is minimal. If we isolate the S1 condition (right part of the Figure) and make another visual assessment, we can see that the vertical axis has a color gradient, which means that in this situation the intelligibility level (the color) changes according to the amount of signal-dependent information. In other words, the result depends strongly on the speaker, which is the goal for patient assessments. The red color, which indicates a functional limitation, appears clearly, meaning that the test can precociously detect a problem.

We set up this situation by using a large set of pseudo-words that respect phonotactic structures frequently found in French. This choice completely neutralizes the abovementioned lexicality or learning effects. Listeners are thus confronted with an acoustic-phonetic decoding task, which mainly involves bottom-up processing and signal-dependent information.

The use of non-words to evaluate speakers with speech production disorders is not new. We can cite for example the work of Shriberg et al. [20] on language and speech disorders in children. The authors propose different metrics based on the binary identification of phonemes. The PPC index is the Percentage of Phonemes Correct [21] identified by listeners on spontaneous speech or on other material as non-words. The binary decision (false/true) is a limitation of the method and authors introduced several alternatives as the PCC-Adjusted where "common clinical consonant distortions are also scored as correct." In order to reduce this binary aspect of identification and in order to analyze the phonetic distortion in a nuanced approach, we have based the comparison of phonemes on the theory of Distinctive Features [22]. In this framework, the phonemes can be decomposed into a set of features which distinguishes them. It is then possible to establish an analog metric where it is possible to distinguish slight

or severe distortions. We hypothesize that such a method will help to obtain sensitive results but also exploitable in terms of typology of distortions.

Our goal is to validate this test on a large population of HNC patients:

1. To ensure that our test truly measures the construct(s) it was designed to measure and that it provides an adequate measure of the theoretical model on which it is based (construct validity). A basic but compelling way to achieve this is to measure the test's ability to separate a control group from a heterogeneously sampled group of patients with disorders ranging from slight to severe.
2. To test the results as compared to a gold standard (concurrent validity), and in particular to go beyond simple binary detection (control vs. patient) and measure the ability of the test to measure the severity of the disorder.

Materials and Methods

This section follows the STARD guidelines (Reporting Diagnostic Accuracy Studies) of the EQUATOR Network [23].

The Design of the Test

The design of the test is described in detail in Lalain et al. [24]. The principle of the test can be summarized as follows. The participants are instructed to pronounce 52 pseudowords drawn randomly from a dictionary containing 89,346 possible forms. The pseudowords share a common structure construed from the isolated elements $C(C)_1V_1C(C)_2V_2$, where V_i are vowels and $C(C)_i$ are either an isolated consonant or a consonant group, as in the forms stoumo, vurtant, muja, charou, leba, ranto, etc. Each list of 52 pseudowords was constructed to be phonetically balanced, which means that every list included, for example {p t k b d g v z ʒ f s ʃ r l m n … } twice in C_1, {pr tr kr gr br fr pl kl…} once in C_1, {a i y u o e ã..} six times in V_1, etc. In a second step, the recordings are transcribed by listeners. As we focused our analysis on the phonological level, these orthographic transcriptions were phonetized and compared to the expected phonetic forms. The result was more sophisticated than a binary decision (correct or incorrect). An algorithm integrating insertion, elision, and phoneme substitution phenomena automatically computed the number of phonological features wrongly decoded by the listener. This score is our measure of intelligibility. One of the advantages of this method is its capacity to provide an analytical analysis of the disorder: our method not only provides a scale of severity (the total score), but it can also point to physiological dimensions such as a predominance of errors in nasality, voicing, or mode or place of articulation.

In this context, if the speaker intends to say something but the listener hears something else, we consider that this constitutes a speech production error because the communication channel is optimal (a silent room, efficient audio playback) and the listener has no hearing impairments.

Table 1. Tumor size and location for the studied population

Tumor localization	Tumor size				Total
	T1	T2	T3	T4	
Tonsil	4	11	4	5	24
Mouth floor	1	4	2	8	15
Root of the tongue	1	5	1	5	12
Oropharynx	0	2	3	2	7
Retromolar	1	3	0	2	6
Tongue	0	4	0	1	5
Mandibula	1	1	0	3	5
Soft palate	2	1	1	0	4
Total	10	31	11	26	78

The foundations of the test are provided in Lalain et al. [24] where we further describe the test design and present preliminary results obtained from 47 speakers. In the current study, we present evidence from HNC speech data obtained from 117 speakers to establish the construct and concurrent validity of our metric.

Patient Group and Control Group

Baseline Demographic

In the framework of the prospective C2SI project [25], we recorded 117 native French speakers (78 patients and 39 healthy subjects) in the oncology rehabilitation unit at the Oncopole in Toulouse, France [25]. Healthy speakers included 21 women (35–76 years old, mean = 59) and 18 men (30–79 years old, mean = 61) who reported the absence of voice/speech disorders. Patients included 35 women (51–87 years old, mean = 66) and 43 men (36–85 years old, mean = 65).

Eligibility Criteria

Patients had to meet the following inclusion criteria:

- Patients with T1 to T4 cancer of the oral cavity and/or oropharynx.
- Patients having received treatment by surgery and/or radiotherapy and/or chemotherapy.
- The recording had to take place at least 6 months after the end of treatment to ensure the stability of the speech disorder, whether audible or not.

Similarly, the criteria for non-inclusion were patients presenting another source of speech impairment (stuttering, for example) or presenting cognitive or visual problems incompatible with the design of the evaluation protocol. These non-inclusion criteria were also used for the recruitment of the control group.

Medical Information

Table 1 shows the number of patients included in the study according to the anatomical region affected by the cancerous lesion and the values of T according to the TNM classification [26] (the internationally accepted standard for cancer staging published by the Union for International Cancer Control).

The most frequent treatment related to the size of the tumors was surgery (84%). The resection of the tumor was associated with a node resection followed in 40% of cases by chemoradiotherapy and in 37% of cases by radiotherapy (RT) only. The delay after the end of the treatment was on average 5 years and 5 months (SD = 55 months).

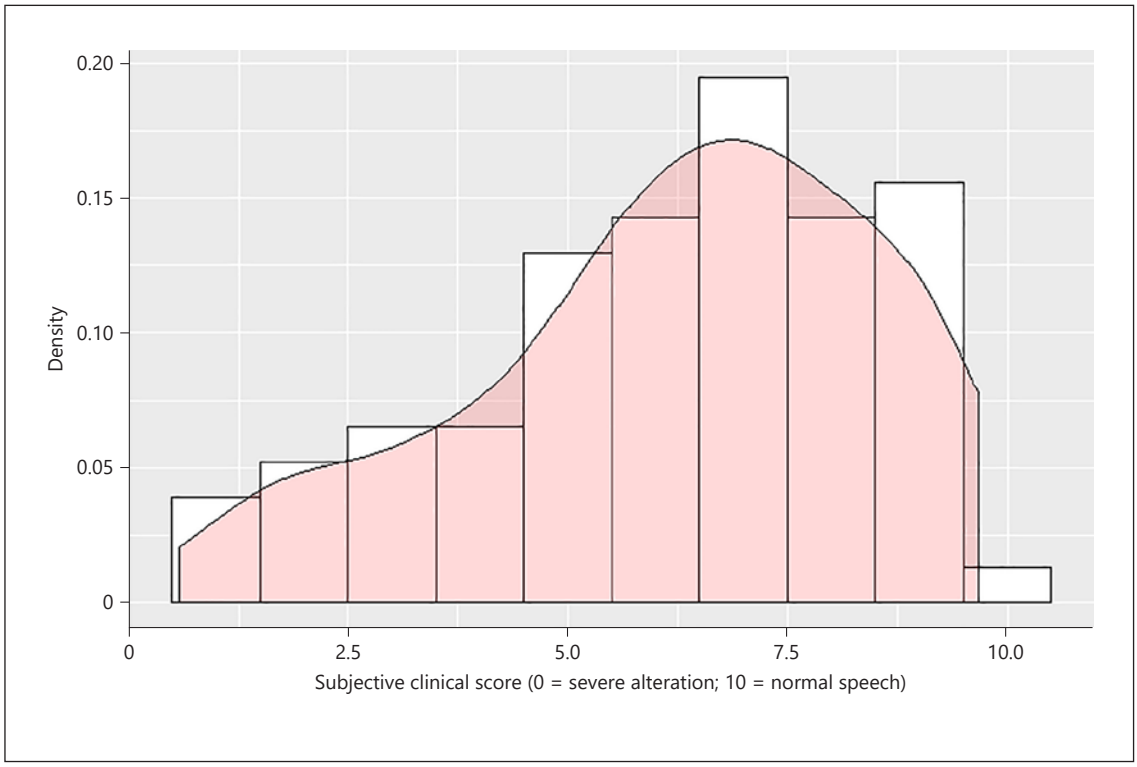


Fig. 2. Distribution of severity of the disease in our population (10 = normal, 0 = severe dysfunction).

Distribution of Severity of the Disorder

All speakers were subjected to an overall assessment of the severity of their disorder using an image description task [25]. This evaluation, described in detail in Balaguer et al. [27], used a visual analog scale from 0 (severe impairment) to 10 (normal speech). The speaker's score was obtained by averaging the scores from 6 speech therapists considered to be experts on speech disorders. An intra-class correlation coefficient was calculated to assess the inter-judge reliability of this subjective evaluation. The high degree of agreement between the jury's scores ($r = 0.77$) demonstrated that the jury was homogeneous and could act as a gold standard [28].

The patients' data showed severity scores between 0.58 (severe impairment) and 9.7 (normal speech) with a mean at 6.2 and a third quartile boundary at 8.0. Twenty-five percent of the patients received scores higher than 8/10, indicating only mild speech impairments (Fig. 2). The group of patients obtained varied scores of disorder severity, which made it possible to verify the metrological strength of our proposal.

Corpus

To record the corpus, the speakers were seated comfortably in an anechoic room in front of a computer screen, which automatically displayed the orthographic form of the pseudoword to be pronounced and produced an audio version at the same time. This double modality, both visual and auditory, was designed to limit reading errors or possible hearing and attentional difficulties. The recordings were carried out with a Neumann TLM 102 cardioid condenser microphone connected to a FOSTEX digital recorder. The sampling frequency was set at 48 kHz.

Each speaker pronounced a different list of pseudowords, drawn randomly from a dictionary of 89,346 possible forms based on identical phonetic constraints. As described in Lalain et al. [24], the same number of phonemes appeared in each list but in different combinations, making the lists equivalent. Once the speaker was recorded, the speech signal was segmented to obtain one audio file per pseudoword. The corpus was therefore composed of 117 speakers × 52 items = 6,084 stimuli.

Perception Task

We recruited 40 native French-speaking everyday listeners with no former experience with a speech disorder and without any hearing impairments to carry out the perception task. Our choice of everyday listeners was based on our desire to create an "ecological" situation in which all listeners are considered specialists in their language and therefore capable of carrying out the task of phonetic decoding without any specific medical or auditory expertise. In other words, this choice was motivated by the fact that the everyday listeners represent "typical communication partners" [29].

Forty listeners transcribed the productions in the corpus using the LANCELOT software program [30]. They received the following instructions: "You will hear a series of non-words. A non-word is a combination of sounds from the French language which has no meaning (e.g., glutu). You will then transcribe what you hear, respecting the rules of French spelling. Certain pronunciations will be difficult to identify, but even in these cases, you will have to provide a transcription."

The stimuli were distributed in different blocks and presented in a random order. Each stimulus was transcribed by 3 different

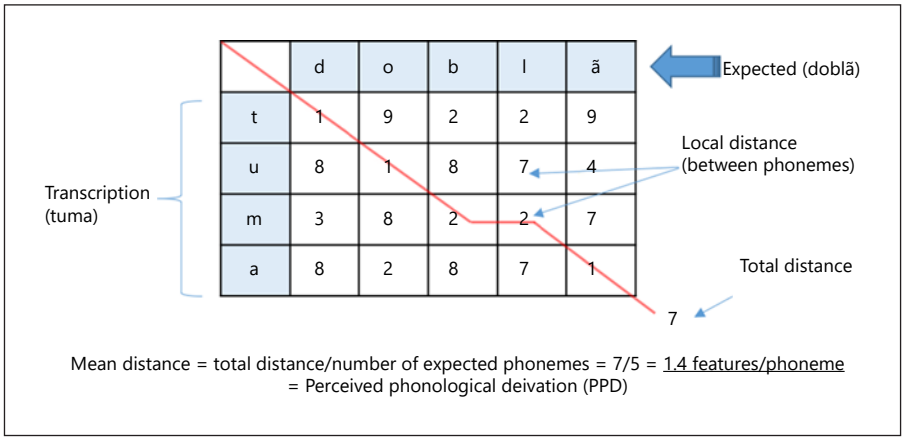


Fig. 3. Comparison of transcribed and expected phonological strings.

listeners, which ultimately represented 18,252 responses (6,084 stimuli × 3). Each listener transcribed about 456 stimuli, which is a part of the productions of each of the 117 speakers. The perception tests took place at the Centre for Speech Experimentation (*Centre d'Expérimentation sur la Parole*, CEP) at the Speech and Language Laboratory (*Laboratoire Parole et Langage,* LPL) in Aix-en-Provence, France. Each listener wore a Superlux HD 681B headset and transcribed the stimuli on a computer. The listeners could adjust the loudness of the audio production to a comfortable level. Each test started with four training stimuli. The items were each presented once automatically, but the listeners could replay them twice more if necessary. The listeners were not subject to any time constraints in performing the test. They themselves chose when to move on to the next item.

Transcription and Response Preprocessing

We collected a total of 18,252 responses from the perception test. As detailed in Ghio et al. [31], these orthographic transcriptions were phoneticized and compared to the expected phonetic forms of the pseudowords using an algorithm based on a calculation of deviant distinctive features between the target form and the transcribed form. This calculation is based on a local distance, which consists in counting the number of distinctive phonological features between two phonemes. For example, the distance between [t] and [d] is equal to 1 (voicing); the distance between [t] and [b] is equal to 2 (voicing, place of articulation); the distance between [b] and [m] is equal to 2 (nasal, sonorant). The final calculation is based on a Wagner-Fischer algorithm that finds the best alignment between the target form and the transcribed form (Fig. 3). This algorithm integrates the phenomena of insertion, elision, and unit substitution. Our measurement, called Perceived Phonological Deviation, or PPD, represents the average number of wrongly perceived features per phoneme. Assuming the communication channel is optimal (a silent room, efficient audio playback) and the listener is a native speaker, the perceived error is therefore directly linked to a production error.

The score was calculated in two steps. First, we determined a score per pseudoword. This involved calculating a score for each pseudoword transcription and then taking the median of the three values, since each pseudoword was transcribed by three different listeners. We thus obtained a consistent score across pseudowords and speakers. In order to ensure the response consistency and the agreement between the listeners, we then applied an outlier detection method. The maximum difference between two phonemes was ten features; a difference in values of half of this distribution (i.e., five features) was considered as acceptable. Accordingly, a result that deviated by ±2.5 features from the median (+ or – 2.5 = 5) was considered as aberrant. For instance, the pseudoword "doba" [doba] was produced by Speaker PRG014 and transcribed as "vovba" [vovba] by Listener 1 (PPD = 0.75), "j'veu'rai" [ʒvøRɛ] by Listener 2 (PPD = 4.25), and "lobaille" [lobaj] by Listener 3 (PPD = 1.25). The median of the PPD is 1.25, whereas the value for Listener 2 (= 4.25) is more than 2.5 features over the median (= 3.75), so we can exclude the second transcription. If a transcription was estimated as aberrant, it was removed from the analysis. We calculated the final score on the average of the remaining values. This procedure allowed us to manage unacceptable inter-listener variability and was especially useful in easily detecting and excluding keyboard typing errors.

As a second step, we calculated the average of the 52 scores per speaker corresponding to the 52 pseudowords produced by the speaker. In sum, we obtained a PPD score for each of the 117 speakers which reflected the average number of deviant features per phoneme, a metric of speech intelligibility. The score reflects an intelligibility loss, where higher scores are associated with reduced intelligibility.

Speaker Classification

In order to measure the ranking ability of the PPD score, we computed a sensitivity/specificity curve used to measure the performance of a binary classifier. The ROC (receiver operating characteristic) function takes the form of a curve which plots the rate of true positives (patients detected as patients) as a function of the rate of false positives (the fraction of healthy subjects who were incorrectly detected as patients) for all classification thresholds [32].

Results

All the statistical tests were carried out using the software environment R, version 3.4.4 [33]. Three sets of results using the PPD score are reported below. These re-

sults illustrate: (i) the PPD score's performance as an outcome variable in an independent mean assessment between the groups (healthy vs. patients); (ii) the score's ranking ability to discriminate between healthy speakers versus patients, and (iii) the correlation strength of the PPD score with an alternative metric (here, a subjective speech severity assessment).

Missing Data and Outliers

Using the filtering technique described in the Transcription and Response Preprocessing section above, we removed 1.2% of the transcripts which mostly contained keyboard typing errors and occasional unacceptable inter-listener variability, where the score for 1 listener was considerably different from the others for the same stimulus.

All of the PPD scores per speaker were retained for analysis; we did not declare any data as an outlier. For graphical convenience, we set the limit of the PPD axes at 3.0 in Figures 4 and 5, although one PPD value was greater than this limit for 1 speaker ($PPD_{BOM94} = 4.07$).

Our data on the correlation between the PPD score and the clinical assessment of disease severity had 15 missing values for the clinical assessment of disease severity: 1 for a patient and 14 for healthy speakers.

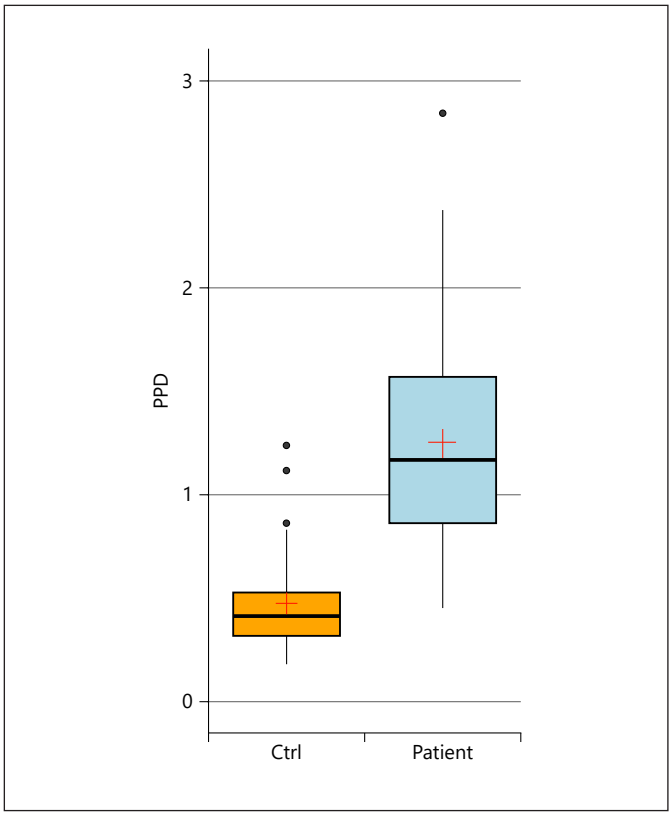
PPD Score as a Function of Group (Healthy Speakers vs. Patients)

Our results revealed that healthy speakers obtained an average PPD score of 0.48 features per phoneme (SD = 0.23; $N = 39$), while patients obtained 1.29 features per phoneme (SD = 0.62; $N = 78$; Fig. 4). Because the PPD data did not display a Gaussian distribution, we performed a logarithmic transformation of the score. We obtained Gaussian distributions (Shapiro test, $p > 0.05$) and homogeneous variances (Bartlett test, $p > 0.05$). In order to determine whether there was statistical evidence that the two populations were significantly different, we ran an independent samples *t*-test with the log-transformed score as a dependent variable and "group of speakers" as a factor. The difference between the two groups was significant ($t(115) = -11.3$; $p < 0.001$).

PPD Score per Speaker and Its Ranking Ability

Figure 5 illustrates the distribution of the PPD score by speaker according to the group (control vs. patients). The horizontal dotted line indicates the optimal patient/control distinction threshold, the calculation of which is explained below.

The ROC curve (Fig. 6) is informative because it predicts classification performance by measuring the area



Color version available online

Fig. 4. PPD score distribution in control subjects (Ctrl) and patients. The mean is marked by a plus sign.

under the curve (AUC). The AUC, which measures the entire two-dimensional area below the whole ROC curve, indicates the probability that the PPD score ranks a patient before a control subject (in the best case, the AUC is equal to 1). The AUC of ROC assesses the diagnostic interest of a test. In our case, the AUC is equal to 0.94, which corresponds to excellent group separability.

The ROC curve also determines the threshold value which will optimize the test. A key question concerns the threshold value of the PPD score, below which all scores should be considered normal and above which all scores should signal a dysfunction. Intuitively, this optimum cut-off point can be identified as being the point on the curve located the furthest from the diagonal line in the ROC curve. Indeed, this point can be computed by the formula: {sensitivity + specificity –1}. When sensitivity = 1 − specificity, the value is equal to 0, it corresponds to a position on the diagonal in the ROC diagram which occurs if the test has no diagnostic value (AUC = 0.5). Conversely, a theoretical value of 1 indicates that there are no false positives or false negatives, i.e., the test is perfect.

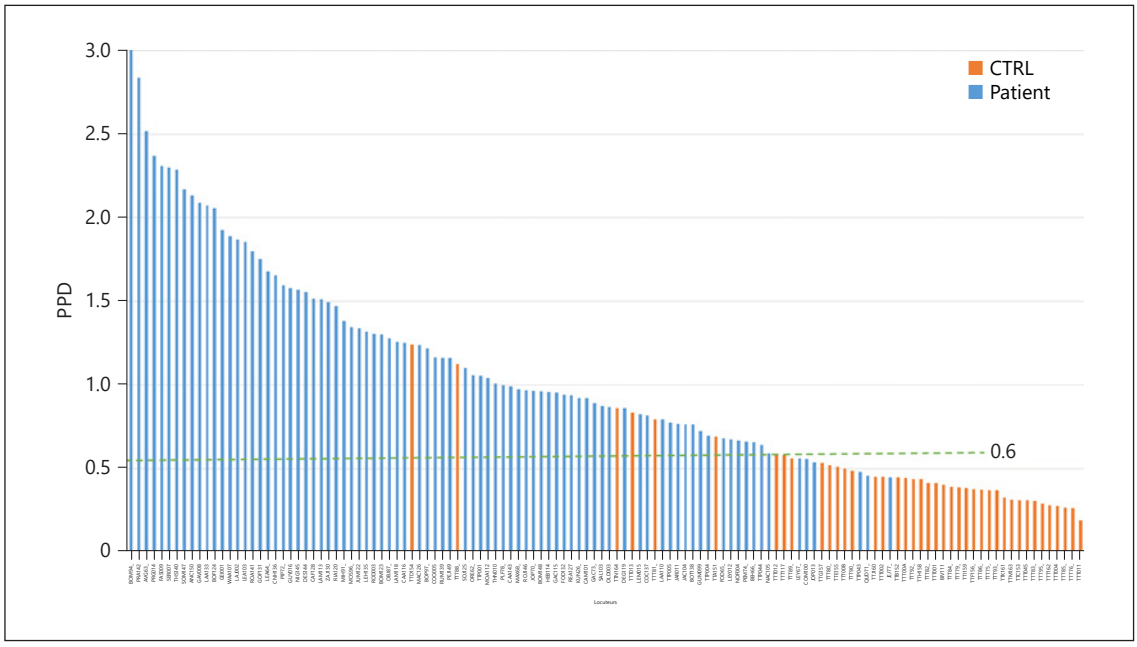


Fig. 5. PPD score per speaker and group.

With real data, it is usual to look for the maximum value of the index, called the Youden Index, in order to use it as a criterion for selecting the optimum cut-off point. In our case, the maximum of the Youden index is 0.783, which corresponds to a PPD threshold equal to 0.6, indicated in Figure 5 by the horizontal dotted line. At this setting, the sensitivity is equal to 0.93 and the specificity is equal to 0.86.

Correlation with the Clinical Measurement of Severity

As described previously, we have a severity index of the disorder for each speaker on an ordinal scale of 0 (severe impairment) to 10 (normal speech). In order to verify the concurrent validity of the PPD test, we examined the correlation between the PPD score and the clinical judgment of severity. These two coefficients were well correlated with an $R_{spearman}$ equal to –0.85.

This correlation was obtained using data from only 102 speakers because the dataset had 15 missing values for the clinical measurement of severity (see Missing Data and Outliers).

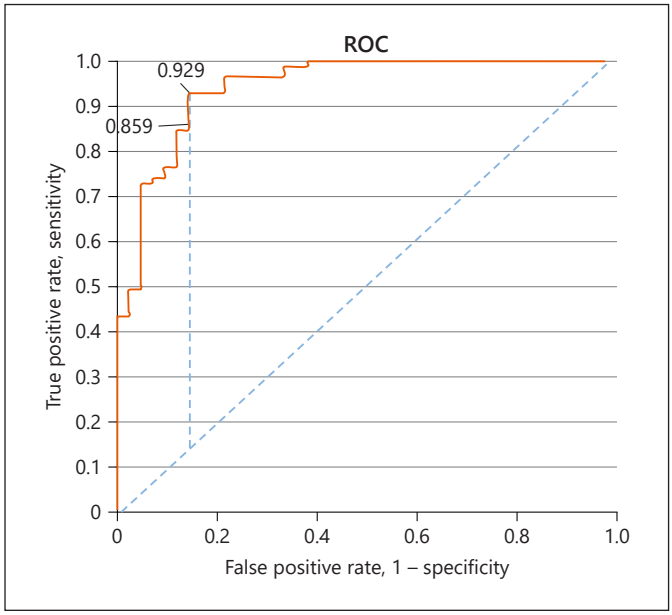
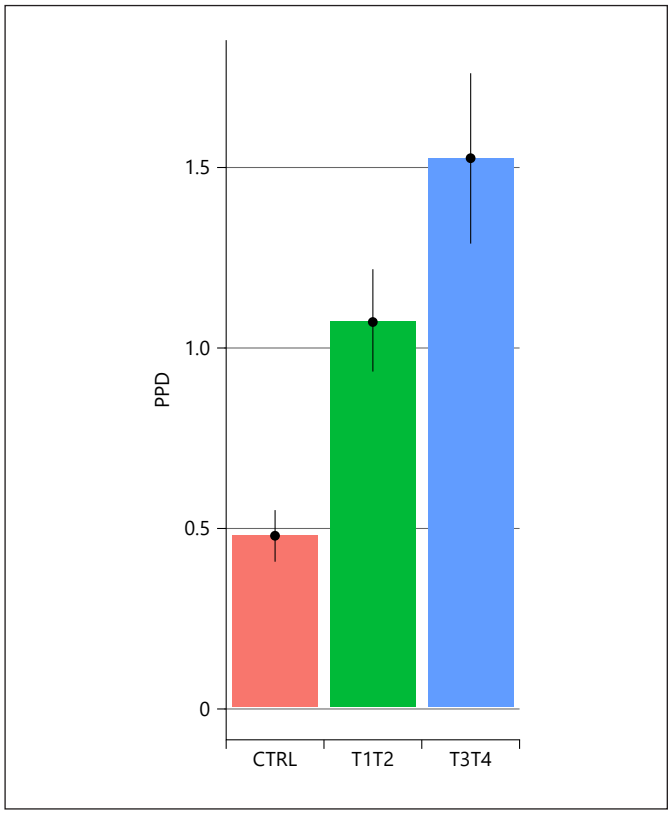


Fig. 6. ROC curve for PPD ranking ability.

Relationship between Tumor Size and PPD Score

It is possible to divide the group of HNC patients into two parts: the T1T2 subset whose tumor size is small and the T3T4 subset whose tumor size is large (see Table 1). Our results revealed that T1T2 patients obtained an average PPD score of 1.07 feature per phoneme (SD = 0.45; N = 41), while T3T4 patients obtained 1.52 feature per phoneme (SD = 0.71; N = 37; Fig. 7). Because the PPD data did not display a Gaussian distribution, we performed a logarithmic transformation of the score. We ob-



Color version available online

Fig. 7. PPD score depending on the size of the tumor. The error bar is the 95% standard confidence interval of the mean.

tained Gaussian distributions (Shapiro test, $p > 0.05$) and homogeneous variances (Bartlett test, $p > 0.05$). In order to determine whether there was statistical evidence that the different groups were significantly different, we ran an ANOVA with the log-transformed score as a dependent variable and "group of speakers" as a three-levels factor (CTRL, T1T2, T3T4). The difference between the groups was significant ($F(2,114) = 76.1$, $p < 0.0001$) and multiple comparison confirmed the differences between all groups, especially between T1T2 and T3T4 ($p = 0.0018$).

Discussion

The results demonstrated strong construct and concurrent validity. The classification strength of our method is excellent in terms of the AUC. Over the threshold of 0.6 features per phoneme, when the measure can be considered to be associated with a dysfunctional limitation, the PPD score is well correlated with the clinical gold standard. This measurement is therefore both capable of discriminating the normal from the pathological and, within the dysfunctional space, accounting for the severity of the disorder.

We are aware that the "patients versus control" comparison is a basic level of construct validity. It is a necessary condition but not sufficient. This validity is confirmed in a second step by the division of the group of patients into coherent subsets at the physiopathological level, taking into account the size of the tumor. The PPD measurement is consistent with the speech production deficit, which increases with the size of the tumor. Another way to further validate the test would be to correlate analytical results with tumor location. For example, damage to the soft palate would be expected to greatly degrade the phonological characteristics linked to nasality, while damage to the tongue could considerably degrade the characteristics linked to the mode and location of articulation. However, these potential correlations can be complicated by other factors: radiotherapy, for instance, can impact various peripheral tissues and make the links between the affected areas and the foreseeable phonetic symptoms opaque. These questions are under study.

The concurrent validity of our metric has been established, since our metric is well correlated with the gold standard of the clinical severity judgment of speech production disorders. Compared to this subjective approach, the task we have designed is a purely linguistic decoding task at a phonological level and not a process of subjective interpretation. Our task therefore corresponds more closely to the typical process of oral communication in which the speaker codes information, which is decoded by the listener. Moreover, as our evaluators were everyday listeners representing "typical communication partners" [29], our approach employs a more typical communication process than traditional clinical evaluation by subjective judgment on a scale. Even if pseudowords are minimalist in terms of communication, they still involve a transfer of information.

Our approach is similar to other techniques seeking to make semantic content unpredictable and to offer non-fixed lists. For instance, in the SUS test [34], a method for the assessment of text-to-speech synthesis intelligibility, sentences are randomly generated, syntactically correct but semantically meaningless, such as: "The table walked through the blue truth," or "How does the day love the bright word?" The authors justify their method as follows: "There are several advantages to this type of material. It controls for the effect of semantic information and this reduction in contextual cues means that ceiling effects are avoided. In addition, because sentences are randomly gen-

erated using a fixed vocabulary, it is possible to generate a very large number of different sentences from the same lists. This reduces the strong learning effects known to occur if sentences are listened to more than once" [34]. We have taken another path, but our approach is ultimately very similar. Due to the linguistic material used (pseudowords), an evaluation using acoustic-phonetic decoding is less dependent on the top-down mechanisms of perception and therefore less dependent on the listener. Independence from the listener reduces the well-known phenomena of variability that weaken the results of perceptual assessments. In addition, using pseudowords has the advantage of giving access to well-mastered, standardized, and very large quantities of linguistic material. Consequently, the PPD score obtained is less subject to evaluation bias.

We reported that only 1.2% of listener responses were excluded from the analyses because they deviated too much from the scores obtained by the other 2 listeners (see Transcription and Response Preprocessing). They corresponded mainly to keyboard typing errors and occasionally to unacceptable levels of inter-listener variability. The volume of discarded data thus remains low and reports to the consistency of our method.

As illustrated in Figure 4, the control subjects exhibited non-zero PPD scores. We can explain this property by the fact that listeners perceive copies of slightly altered phonemes, including by non-pathological speakers. Here, we measured "non-pathological" distortions that occur in speech production and that are generally rectified by the listener by accessing the lexicon and meaning. Because our test is based on a single acoustic-phonetic decoding task, these mechanisms were – as expected – well inhibited and therefore did not allow for any restoration. These observations testify to the sensitivity of the test to capture even very small speech disturbances. Such a capacity is important for detecting early signs of dysfunction, unlike conventional tests, which can only measure severe degradation.

Figure 5 illustrates a fairly distinct distribution between healthy subjects (low scores) and the majority of patients (higher scores), which shows that our test provides an accurate ability to discriminate between the two groups. Some patients received low scores, which may reflect a low functional impact of cancer treatment on their speech. Conversely, some speakers in the control group were set apart by their high PPD scores. A closer analysis of these productions reveals a particular difficulty or lack of attention in producing pseudowords. For example, one of the speakers pronounced the pseudoword "minso" /mɛ̃so/ as [mjozo] and the sequence "plouco" /pluko/ as [plɔkso]. The listener's transcription of these words is therefore correct, reflecting productions that deviated from the targets "minso" and "plouco" and generating higher PPD values. This example highlights one of the limits of our method, which is that it cannot be applied in cases involving subjects with reading or phonological difficulties that cause an inability to produce pseudowords correctly. We also note the importance of attentional processes needed to read and then produce pseudowords; our subjects could not benefit from the facilitation that they usually encounter with frequent words.

The concurrent validity of our method can be improved with other assessments studied in the framework of the C2SI project [25]. For example, we plan to compare PPD scores of patients with a Sentence Verification Task (SVT) [35], which is a method oriented toward comprehensibility [36]. This comparison would make it possible to distinguish the functional limitations measured by the PPD score and the impact of the disability on understanding the speaker's message. A difference between these two elements could be interpreted as a good use of adjustment phenomena. If speakers with high PPD are well understood by the SVT test, this could indicate their capability to overcome their functional limitations and to reach their communicational goals.

Within the framework of the ANR-18-CE45-0008 Rugbi Project (https://anr.fr/Project-ANR-18-CE45-0008), we are also studying acoustic-phonetic decoding by examining the effect of linguistic factors on the perceptual identification of intervocalic consonants in a reading task. The PPD score used was computed on single consonants taken from the continuous speech recorded in a reading task. The results are being examined as a function of consonant nature, oral/nasal vocalic context, word class (function or content), and prosodic position within sentences [37]. Within the same project, we are also comparing the PPD score with the results of automatic speech analysis techniques as described in Laaridh et al. [38] and Abderrazek et al. [39], which will contribute to the test's validity.

In terms of implications for clinical practice, we are developing an application that will allow our test to be administered during a consultation. The software requires two screens, an audio headset, and a microphone. Pseudowords are displayed on a screen visible only to the patient and are played back to the patient in their audio form through headphones. The patient then produces the pseudoword, and the evaluator transcribes what they hear. The software records the written response and compares it to the target. The patient's speech signal is also recorded for archiving or further analysis. At the end of the exam, the PPD score is calculated automatically based

on the answers. In order to reduce the length of the test, we are working on optimizing the linguistic material used in the test by identifying an optimal tradeoff between its effectiveness and efficiency [40]. This study also validates the list equivalence: the same speaker, tested with two phonetically balanced but randomly generated lists, obtained identical results on each list. A number of new research avenues have emerged from these results. For instance, calculating the confusion matrices of phonemes would make it possible to go beyond the scalar value of the PPD score. A more detailed analysis of the altered features could be of great value for therapeutic orientation.

Turning to clinical applications other than cancer-related speech disorders, there are at least several acquired and developmental speech disorders that could benefit from the pseudoword-based intelligibility tool we propose here. In general terms, the PPD score focuses on feature-based intelligibility, that is, speech disturbances at the segmental level. Such disturbances are observable in several neurologically based speech impairments such as dysarthrias. Slurred speech and speech sound misarticulations in particular are a hallmark of all dysarthric types [28]. Because the PPD test appears to be sensitive to minor speech disturbances, it could also be a valuable tool to reliably detect early speech disturbances in non-fluent variants of primary progressive aphasias, as well as for differential diagnosis to discriminate between non-fluent and logopenic variants of primary progressive aphasias [41]. In contrast, the PPD intelligibility score would be inappropriate to capture concomitant aspects of the abovementioned speech disorders, such as suprasegmental variations, fluency and voice impairments, or higher-level language disorders (notably in aphasia), which would call for other assessment tools and materials.

It could be argued that non-word-based materials for intelligibility assessment introduce a bias related to the fact that non-word processing require strong phonological memory skills and phonemic awareness [42], abilities that are problematic in a number of speech-language disorders. Non-word repetition would thus put an additional difficulty for language-impaired speakers and the final score would reflect that added difficulty. However, reduced intelligibility does not exclusively reflect poor articulation skills, in fact, it may arise at different levels in the process of word-form encoding, including phonological disorders. A common tool to assess speech intelligibility in complex multidimensional impairments is crucial. It is the analysis of error patterns elicited with phonetically balanced lists that will give further insights into the source of intelligibility reduction.

Conclusion

Our new metric for testing intelligibility, defined as the amount of speech understood only from signal-dependent information, is designed to overcome the limits of traditional intelligibility tests. To do so, we used lists of pseudowords extracted from a directory of tens of thousands of elements. Our evaluators were everyday listeners who represented "typical communication partners" [29]. Our method was validated by construct and concurrent validity on a population of 117 speakers: 39 healthy subjects and 78 patients with cancer of the oral cavity and the oropharynx. Patients compared to the control group demonstrated significantly higher PPD scores, indicating higher numbers of deviations. If we use the measure as a predictor to identify patients versus control group subjects, the AUC of the ROC curve corresponds to excellent group separability. The threshold of 0.6 altered features per phoneme appears to be the limit between healthy and pathological speech in this task. Finally, the analysis showed a close correlation between the PPD and a clinical judgment of the disorder severity obtained from experts. As only 1.2% of the answers were rejected, we can conclude that the test is efficient, despite certain limits, including the involvement of subjects with reading difficulties or attentional deficits. The most important advantage of the test is to detect subtle differences in intelligibility, avoiding the ceiling effect that can be found in traditional intelligibility tests. The PPD measurement does not suffice on its own to understand all the characteristics of a dysfunctional speaker, nor to fully characterize their disability to communicate naturally, but our method could be an important addition to other tools in this aim. The PPD metric may also be well suited for assessments of other pathologies. Our method is currently applicable in clinical settings, though testing with other data will make it possible to further validate our method in the service of patients with speech disorders.

Acknowledgements

The authors thank the CEP staff (www.lpl-aix.fr/~cep), especially Carine André, for assisting in the perceptual experiments. The authors thank Oriana Reid Collins (Collins Traduction, https://collinstraduction.com/) for her English proofreading work in the text.

Statement of Ethics

The clinical study was conducted at the Toulouse Oncopole in accordance with the World Medical Association Declaration of Helsinki. A favorable decision was obtained from the research ethics committee of the Toulouse hospitals on May 17, 2016. All participants gave their written informed consent. To guarantee anonymity, a code was assigned to each participant. A declaration was made concerning data processing to the *Commission nationale de l'informatique et des libertés* (the National Data Protection Authority, No. 1876994v 0, July 24, 2015).

Conflict of Interest Statement

The Authors declare that there are no conflicts of interest.

Funding Sources

This study was supported by grant No. 2014-135 from the French National Cancer Institute (INCa) C2SI project and by the grant ANR-18-CE45–0008 "Looking for relevant linguistic units to improve the intelligibility measurement of speech production disorders" from the National Research Agency. The study also benefitted from a doctoral contract (Marie Rebourg) supported by the French National Cancer Institute.

Author Contributions

Alain Ghio and Muriel Lalain conceived and planned the experiments. Virginie Woisard selected the patients and recorded the sound data. Alain Ghio, Muriel Lalain, and Marie Rebourg carried out the experiments. Alain Ghio, Anna Marczyk, and Corinne Fredouille processed the data. Virginie Woisard supervised the project. All authors contributed to the interpretation of the results. Alain Ghio took the lead in writing the manuscript. All authors provided critical feedback and helped shape the research, analysis, and manuscript. All authors approved the final version for publication. All authors agree to be responsible for all aspects of the work, ensuring that questions relating to the accuracy or integrity of any part of the work are properly investigated and resolved.

Data Availability Statement

All data analyzed during this study will be available at the end of the ANR-18-CE45-0008 RUGBI Project (https://anr.fr/Project-ANR-18-CE45-0008) in 2023. Further enquiries can be directed to the corresponding author.

References

1 Meyer TK, Kuhn JC, Campbell BH, Marbella AM, Myers KB, Layde PM. Speech intelligibility and quality of life in head and neck cancer survivors. Laryngoscope. 2004;114(11):1977–81.
2 Karnell LH, Christensen AJ, Rosenthal EL, Magnuson JS, Funk GF. Influence of social support on health-related quality of life outcomes in head and neck cancer. Head Neck. 2007;29(2):143–6.
3 Badley EM. An introduction to the concepts and classifications of the international classification of impairments, disabilities, and handicaps. Disabil Rehabil. 1993;15(4):161–78.
4 Enderby P. Frenchay dysarthria assessment. Int J Lang Commun Disord. 1980;15(3):165–73.
5 Yorkston KM, Strand EA, Kennedy MRT. Comprehensibility of dysarthric speech. Am J Speech Lang Pathol. 1996;5(1):55–66.
6 Lindblom B. On the communication process: speaker-listener interaction and the development of speech. Augment Alt Commun. 1990;6(4):220–30.
7 Fontan L, Tardieu J, Gaillard P, Woisard V, Ruiz R. Relationship between speech intelligibility and speech comprehension in babble noise. J Speech Lang Hear Res. 2015;58(3):977–86.
8 Kent RD, Weismer G, Kent JF, Rosenbek JC. Toward phonetic intelligibility testing in dysarthria. J Speech Hear Disord. 1989 Nov;54(4):482–99.
9 Weismer G. Speech intelligibility. In: Ball MJ, et al., editors. The handbook of clinical linguistics. Wiley; 2008. p. 568–82.
10 Tikofsky RS. A revised list for the estimation of dysarthric single word intelligibility. J Speech Hear Res. 1970;13(1):59–64.
11 Yorkston KM, Beukelman DR. Communication efficiency of dysarthric speakers as measured by sentence intelligibility and speaking rate. J Speech Hear Disord. 1981;46(3):29–301.
12 Yorkston K, Beukelman DR, Tice R. Sentence intelligibility test [Measurement instrument]. Lincoln: Tice Technologies; 1996.
13 Enderby PM, Palmer R. FDA-2: Frenchay Dysarthria Assessment: Examiner's Manual. Pro-Ed; 2008.
14 Warren RM, Warren RP. Auditory illusions and confusions. Sci Am. 1970 Dec;223(6):30–6.
15 Ganong WF 3rd. Phonetic categorization in auditory word perception. J Exp Psychol Hum Percept Perform. 1980 Feb;6(1):110–25.
16 Samuel AG. Phonemic restoration: insights from a new methodology. J Exp Psychol Gen. 1981 Dec;110(4):474–94.
17 Auzou P, Rolland-Monnoury V. Batterie d'Evaluation Clinique de la Dysarthrie. Ortho-édition; 2006.
18 Ghio A, Giusti L, Blanc E, Pinto S. French adaptation of the "Frenchay Dysarthria Assessment 2" speech intelligibility test. Eur Ann Otorhinolaryngol Head Neck Dis. 2020 Mar;137(2):111–6.
19 Rebourg M, Lalain M, Ghio A, Fredouille C, Fakhry N, Woisard V. Learning effect on words list during intelligibility assessment: an alternative with pseudowords. Clin Linguist Phon. submitted.
20 Shriberg LD, Lohmeier HL, Campbell TF, Dollaghan CA, Green JR, Moore CA. A nonword repetition task for speakers with misarticulations: the Syllable Repetition Task (SRT). J Speech Lang Hear Res. 2009 Oct;52(5):1189–212.
21 Shriberg LD, Austin D, Lewis BA, McSweeny JL, Wilson DL. The percentage of consonants correct (PCC) metric: extensions and reliability data. J Speech Lang Hear Res. 1997 Aug;40(4):708–22.
22 Jakobson R, Fant G, Halle M. Preliminaries to speech analysis: the distinctive features and their correlates. MIT Press; 1951.
23 Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015 Oct;351:h5527.
24 Lalain M, Ghio A, Giusti L, Robert D, Fredouille C, Woisard V. Design and development of a speech intelligibility test based on pseudowords in French : why and how? J Speech Lang Hear Res. 2020 Jul;63(7):2070–83.

25 Woisard V, Astésano C, Balaguer M, Farinas J, Fredouille C, Gaillard P, et al. C2SI corpus: a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers. Language Resources and Evaluation. Berlin: Springer; 2020.
26 Brierley JD, Gospodarowicz MK, Wittekind C. TNM classification of malignant tumours. John Wiley & Sons; 2016.
27 Balaguer M, Boisguérin A, Galtier A, Gaillard N, Puech M, Woisard V. Assessment of impairment of intelligibility and of speech signal after oral cavity and oropharynx cancer. Eur Ann Otorhinolaryngol Head Neck Dis. 2019 Oct;136(5):355–9.
28 Duffy JR. Motor speech disorders: Substrates, differential diagnosis, and management. St. Louis: Elsevier; 2013.
29 Nagle KF, Eadie TL, Yorkston KM. Everyday listeners' impressions of speech produced by individuals with adductor spasmodic dysphonia. J Commun Disord. 2015;58:1–13. https://doi.org/https://doi.org/10.1016/j.jcomdis.2015.07.001.
30 André C, Ghio A, Cavé C, Teston B. PERCEVAL : A computer-driven system for experimentation on auditory and visual perception. International Congress of Phonetic Sciences (ICPhS), Barcelona, 2003, 1421–4. https://hal.archives-ouvertes.fr/hal-00142980.
31 Ghio A, Lalain M, Giusti L, Fredouille C, Woisard V. How to compare automatically two phonological strings: application to intelligibility measurement in the case of atypical speech. 12th Conference on Language Resources and Evaluation, LREC, 2020, 1682–7. https://hal.archives-ouvertes.fr/hal-02482615.
32 Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem. 1993 Apr;39(4):561–77.
33 R Core Team. R: a language and environment for statistical computing. Vienna; R Foundation for Statistical Computing; 2017. http://www.R-project.org/.
34 Benoît C, Grice M, Hazan V. The SUS test : a method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. Speech Commun. 1996;18(4):381–92.
35 Pisoni DB, Manous LM, Dedina MJ. Comprehension of natural and synthetic speech: effects of predictability on the verification of sentences controlled for intelligibility. Comput Speech Lang. 1987 Sep;2(3-4):303–20.
36 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. XXXIIe Journées d'Études Sur La Parole. 2018. doi: 10.21437/JEP.2018-23.
37 Duez D, Ghio A, Viallet F. Effect of linguistic context on the perception of consonants in Parkinsonian Read French speech. Clin Linguist Phon. 2020 Oct;0(0):1–19.
38 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. Proc Interspeech. 2018;1266:2943–7. https://hal.archives-ouvertes.fr/hal-01962170 https://doi.org/10.21437/Interspeech.2018-1266.
39 Abderrazek S, Fredouille C, Ghio A, Lalain M, Meunier C, Woisard V. Towards interpreting deep learning models to understand loss of speech intelligibility in speech disorders — step 1: CNN model-based phone classification. Proc Interspeech. 2020;2239:2522–6.
40 Marczyk A, Ghio A, Lalain M, Rebourg M, Fredouille C, Woisard V (2021) Optimizing linguistic materials for feature-based intelligibility assessment in speech impairments. Behav Res. 2021. doi: https://doi.org/10.3758/s13428-021-01610-9.
41 Ogar JM, Dronkers NF, Brambati SM, Miller BL, Gorno-Tempini ML. Progressive nonfluent aphasia and its characteristic motor speech deficits. Alzheimer Dis Assoc Disord. 2007 Oct-Dec;21(4):S23–30.
42 Clark NB, McRoberts GW, Van Dyke JA, Shankweiler DP, Braze D. Immediate memory for pseudowords and phonological awareness are associated in adults and pre-reading children. Clin Linguist Phon. 2012 Jul;26(7):577–96.