Behavior Research Methods (2022) 54:42–53
https://doi.org/10.3758/s13428-021-01610-9

Optimizing linguistic materials for feature-based intelligibility assessment in speech impairments

A. Marczyk[1] · A. Ghio[1] · M. Lalain[1] · M. Rebourg[1] · C. Fredouille[2] · V. Woisard[3,4]

Accepted: 2 May 2021 / Published online: 7 June 2021

Abstract
Assessing the intelligibility of speech-disordered individuals generally involves asking them to read aloud texts such as word lists, a procedure that can be time-consuming if the materials are lengthy. This paper seeks to optimize such elicitation materials by identifying an optimal trade-off between the quantity of material needed for assessment purposes and its capacity to elicit a robust intelligibility metrics. More specifically, it investigates the effect of reducing the number of pseudowords used in a phonetic-acoustic decoding task in a speech-impaired population in terms of the subsequent impact on the intelligibility classifier as quantified by accuracy indexes (AUC of ROC, Balanced Accuracy index and F-scores). A comparison of obtained accuracy indexes shows that when reduction of the amount of elicitation material is based on a phonetic criterion—here, related to phonotactic complexity—the classifier has a higher classifying ability than when the material is arbitrarily reduced. Crucially, downsizing the material to about 30% of the original dataset does not diminish the classifier's performance nor affect its stability. This result is of significant interest to clinicians as well as patients since it validates a tool that is both reliable and efficient.

Keywords intelligibility · speech impairments · speech material · data reduction

Introduction

Intelligibility measurement is one of the main clinical instruments by which speech impairments are assessed. As a diagnosis, intervention and monitoring tool it serves multiple purposes in the clinical setting. Since typically clinical decisions are partly based upon assessments of patient intelligibility, it is crucial that the metrics that any instrument provides are reliable and objective. The reliability of an intelligibility index will depend on various variables related to the linguistic materials used to elicit spoken output, the elicitation procedure, the individual listener, and the exact nature of the impairment the instrument is intended to evaluate, among other things. This paper is concerned with the first of these factors—the linguistic materials—and examines how such materials can be optimized so that they provide a robust and stable estimate of intelligibility loss. Although the materials described here, which focus on the feature-level assessment of impaired speech, that is, atypicalities in the articulation of consonants and vowels, are primarily intended for use with speech impairments arising specifically from speech sequelae of head and neck cancer (henceforth HNC), they can be equally applicable to a wide range of speech impairments characterized by segment distortions such as dysarthria (e.g., Kent et al., 1989) or apraxia of speech (e.g., Haley et al., 2017).

The development of elicitation materials focusing on the segmental speech dimension for the purpose of eliciting intelligibility scores has been guided by two important findings. First, it has been shown that carefully structured elicitation materials are probably better suited to quantifying reduced intelligibility than materials where linguistic factors are not controlled for (e.g., Kent et al., 1989; Kent & Kim, 2011; Miller, 2013). Specifically, phonetically balanced word lists have been shown to be more sensitive to capture subtle articulatory distortions, thus making it possible to directly interpret scores in phonetically meaningful terms. Speech elicited using such lists not only is helpful for classification purposes (i.e., to carry out a binary discrimination between healthy and speech-impaired speech) but also offers information about the underlying speech deficit and pinpoints its precise locus.

✉ A. Marczyk
ana-katarzyna.MARCZYK-BUKLAHA@univ-amu.fr

[1] CNRS, LPL, Aix-en-Provence, Aix Marseille Univ, 5 Avenue Pasteur, 13100 Aix-en-Provence, France

[2] LIA, Avignon University, Avignon, France

[3] Service ORL, CHU Larrey, Toulouse, France

[4] URI Octogone-Lordat, Toulouse, France

Nonetheless, using word-based elicitation materials has important caveats. Top-down lexical effects can promote phonemic restoration in the listener (Samuel, 1981), while repeated exposure to the same word stimuli can induce familiarization effects (Lagerberg et al., 2015), both issues that are important sources of bias in intelligibility assessment. To circumvent such interferences, phonetically motivated lists of pseudowords have been proposed (Allen et al., 2012; Barreto et al., 2010). One such list has been recently designed with the aim of eliciting an intelligibility index labelled Perceived Phonological Deviation (henceforth PPD, Lalain et al., 2020) for French-speaking speech-disordered populations. It consists of a corpus of 90,000 phonetically controlled pseudowords from which it is possible to generate different but equivalent lists of 52 forms. With regard to the first of the concerns noted above, a recent study in which intelligibility was rated using the PPD score showed no familiarization effects in speakers when they read lists based on the pseudoword corpus as compared to word lists extracted from the BECD dysarthria assessment battery (Rebourg et al., 2020).

Besides the type of linguistic materials used to elicit speech production for intelligibility assessment, a second issue is related to the trade-off between the instrument's ability to judge intelligibility and the volume of data needed to generate statistically reliable results. Clinical practice imposes considerable time constraints on practitioners and obtaining a reliable yet efficient speech performance classifier is of paramount importance. While PPD intelligibility scoring has proved to be highly effective at discriminating between healthy and speech-impaired speakers, having to elicit 52 pseudowords requires a considerable amount of time on the part of clinicians. By way of comparison, the BECD dysarthria assessment battery (Enderby, 1983), whose French adaptation (Auzou & Rolland-Monnoury, 2006) is widely used by French-speaking speech therapy practitioners, generates intelligibility scores from ten words and ten sentences, randomly selected from lists of 50 items, and thus requires less time for clinicians to perform. A short testing time is equally important for patients, given that fatigue often causes patients to leave the task incomplete.

Thus, the general objective of this paper is to optimize the linguistic elicitation materials used in intelligibility assessment by identifying an optimal trade-off between effectiveness and efficiency, measured as a function of three parameters: the accuracy of the instrument as an intelligibility classifying tool, the sample size needed to obtain this accuracy and the stability of the result. The paper is structured around three experiments based on the above-mentioned pseudoword materials used to elicit PPD scores. Experiment 1 tackles the problem of variability in estimated intelligibility due to lack of equivalence between lists. Experiment 2 examines how the sample size influences the predictive accuracy of the PPD score to discriminate between groups. Finally, experiment 3 focuses on the effect of the material's phonetic content on the ability of the instrument to classify the input. The main hypothesis under scrutiny is that phonetically guided data reduction is the best way to meet the optimization criteria.

Methods

The assessment materials described in this paper is part of the Cancer-Related Speech Severity Index (CS2I) project, whose aim is to measure the impact of speech treatments for cancers of the oral cavity and oropharynx (i.e., head and neck cancers) using a combination of automatic assessment performed by software and perceptual assessment performed by clinicians (Astésano et al., 2018). The research protocol was reviewed by the University Hospital Centre of Toulouse's Research Ethics Committee (CER), which analyses the ethical aspects of research protocols directly or indirectly involving humans. Following CER approval on 17 May 2016, the CS2I project was registered with the French government's Commission Nationale de l'Informatique et des Libertés (CNIL) on 24 July 2015 under number 1876994v0. As noted, the stipulated goal of the project is to record the speech of patients treated for HNC cancers, and thus far to this end 85 cancer patients and 41 healthy speakers have been recorded in a battery of linguistic and cognitive tasks, prior to which all participants or their legal guardians signed an informed consent.

Participants

The experiments reported in this paper include healthy and speech-impaired speakers recovering from cancer of the oral cavity or oropharynx (HNC).

All speech-impaired samples were recorded in the Oncology Rehabilitation Centre at the Oncopole Institute in Toulouse. These speakers were patients who had received treatment (surgery, radiotherapy, or chemotherapy) following a T1-T4 cancer of either the oropharynx or oral cavity. All patients were recorded at least 6 months after treatment to ensure the stability of the speech deficit, independently of whether their speech was perceived as distorted or not. Concomitant speech disorders such as stuttering and cognitive or visual deficits constituted exclusion criteria. Likewise, healthy speakers recruited as controls for the experiments reported no speech, hearing or visual impairments and were matched in age, sex, and socio-educational background. Table 1 summarizes clinical and control group characteristics.

We will refer to the group recruited for the list equivalence experiment (experiment 1) as cohort 1. It was made up of ten healthy and ten speech-impaired speakers. We will call the participants in the sample size experiment (experiments 2

and 3) cohort 2. It included 126 speakers (41 healthy subjects and 85 patients).

Elicitation materials

As mentioned in the Introduction, all three experiments employed pseudoword materials described in Lalain et al. (2020). The short description of these materials provided below will be sufficient to contextualize our research. Further details are available in the original paper.

The elicitation materials used for the task consisted of 52 disyllabic pseudowords characterized by the same phonotactic structure C_1 V_1 C_2 V_2, where V_1 and V_2 correspond to single vowels and C_1 and C_2 correspond to either a single consonant or a consonant cluster. C_1 and C_2 represent the most frequent singletons and consonant clusters in French, accounting for at least 87% of all produced consonants at each phonetic position (that is, initial and intervocalic). Possible combinations between them allow the generation of 90,000 pseudowords (after exclusion of semantically meaningful items), a database from which equivalent pseudoword lists are generated. Table 2 provides a summary of the consonants and vowels used in the pseudoword corpus. The number of pseudowords in the final list (52) is phonetically motivated and intended to ensure the high robustness of the proposed metric, robustness referring in this case to the fact that it is possible to obtain multiple samples of each speech sound so that any subset of the 52-item list is equally representative of the French sound system. Specifically, each consonant appears at least twice in each position, that is, once as a singleton and once in a consonant cluster, while each vowel appears at least six times in each syllable. Because several singletons (such as / / or /∫/) do not frequently form clusters with another consonant, single consonants are set up to come out twice each in C_1 and C_2 positions.

Speech samples were recorded as follows. Speakers were comfortably installed in an anechoic room in front of a computer screen. To avoid errors due to reading, hearing or attentional difficulties, the target pseudoword was displayed simultaneously in its visual (i.e., orthographic) and auditory form (see Astésano et al., 2018 for details). The recordings were made with a Neumann TLM 102 cardioid condenser

Table 1 Summary of patient and control speaker characteristics

Characteristics	Patients (n = 85)	Control group (n = 41)
Mean age	65 ± 9	60 ± 13
Men:women	47:38	17:24
Severity assessment*	6.15 ± 2.28	
Tumor region	oral cavity 33; 52 oropharynx	
Tumor localization	tonsil:25; base of the tongue: 15; tongue: 7; mandible: 5; oropharynx: 8; floor of the mouth: 15; retromolar gap: 6; velum: 4	
Tumor size	size 1:10; size 2:35; size 3:13; size 4:27	
Tumor malignancy	T0:23; T1:19; T2a: 7; T2b:13; T2c: 4; T3: 5; not informed: 14	
Anatomic pathology	adenoid cystic:5; squamous cell carcinoma:80	
Time post treatment (months)	63 ± 54	
Initial or recurrent surgical treatment yes:no	71:14	
Surgical treatment involving the lymph node yes: no	74:11	
Reconstruction yes:no	57:25	
Radiotherapy yes:no	80:5	
Chemotherapy yes:no	45:40	
Recurrence yes:no	26:59	
Physical pain score**	62 ± 25	
General health score	56 ± 17	
Vitality score	52 ± 20	
Functioning and social welfare score	64 ± 24	
Mental health score	62 ± 21	

*Severity scores obtained independently from six speech therapists reflect their averaged subjective perception of patients' speech and range from 0 (severe impairment) to 10 (normal).

**Scores extracted from a patient-reported health survey MOS SF-36 (Leplège et al., 2001). Varying from 0 to 100, lower scores reflect self-perceived poor health, loss of function, and the presence of pain.

Table 2 Summary of vowels and consonants that can appear in each of the phonetic contexts in the pseudowords used in these studies

Position	Segmental content
C_1 = singleton	p t k b d g v z f s ʃ l m n ŋ j
C_1 = cluster	p t k g b f pl kl fl st bl sk sp gl d ps
V_1	a i y u O* E ɑ̃ ɛ̃
C_2 = singleton	p t k b d g v z f s ʃ l m n ŋ j
C_2 = cluster	st ks d s kt n pl g d kl j lt v v gz p t t bl m p k sk b sp k f fl b gl ps pt
V_2	a i y u O E ɑ̃ ɛ̃

* Capital letters represent archiphonemes, that is, a class of phonemes that share all but one feature (here, vowel height).

microphone connected to a FOSTEX digital recorder. The sampling frequency was set at 48 kHz. The recording session took about 2–3 min.

Each speaker read one list (cohort 2, experiments 2 and 3) or two lists (cohort 1, experiment 1) of 52 pseudowords, randomly generated from the pseudoword corpus. The audio recordings were then segmented and each pseudoword was saved as a separate audio file.

Acoustic-phonetic decoding

Eighteen transcribers were recruited to phonetically and acoustically code speaker output from the list equivalence experiment and 40 transcribers did the same for the output resulting from the sample size experiment. All transcribers were native French speakers with no self-reported hearing deficits. Because we were interested in an ecological assessment, without possible bias due to clinical expertise, we selected only naïve listeners with no prior speech pathology experience. The listeners transcribed the productions of the corpus using PERCEVAL software (André et al., 2003). They received the following instructions: "You will hear pseudowords. A pseudoword is a combination of sounds of the French language which has no meaning (e.g., "glutu"). Respecting the rules of spelling in French, transcribe what you hear. Certain pronunciations may be difficult to identify, but in every instance, you must provide a transcription."

The recordings of the pseudowords were distributed in several blocks and presented in random order in terms of item and healthy or speech-impaired speaker. Each recorded word was transcribed by three different listeners. The coding procedure took place at the Speech Experimentation Centre (http://cep.lpl-aix.fr/) at the Laboratoire Parole et Langage in Aix-en-Provence, France. Working individually, each listener wore Superlux HD 681B headphones and set their own playback volume level Before they began to transcribe participant output, the transcribers heard four training words in order to familiarize themselves with the procedure. Thereafter, each item to be transcribed was presented once, although the listener could repeat the playback twice.

Intelligibility scoring

First, all of the pseudowords used in the experiment materials were phonetically transcribed using LIA_PHON software, a French text-to-phoneme converter (Bechet, 2001), and the same procedure was then applied to the playback-based transcriptions. Then an intelligibility score was computed for each recorded item using the Wagner–Fisher algorithm by comparing the expected (list-based) and actual (recording-based) transcriptions (Ghio et al., 2020). The score expressed the degree of dissimilarity (i.e., deviation), calculated in terms of distinctive features (maximum six), between 35 French phonemes retained for the protocol. For example, if the recording-based transcription differed from the list-based transcript in three phonological features, this yielded a PPD score of 3. The higher the PPD score, the greater the distance between the expected and actual transcriptions and therefore the greater the loss of intelligibility. A final PPD score was computed for each speaker by averaging their PPD scores across pseudowords.

Data analyses

All subsequent analyses were performed using R language (R Development Core Team, 2013) with the R Studio interface (RStudio Team, 2015) using customized packages and in-house scripts. The significance level was established at $\alpha = 0.05$.

List consistency in experiment 1 was evaluated by means of Pearson correlation analyses for bivariate normally distributed data (Mardia skewness = 8.38, $p = .08$, Mardia kurtosis = 0.39, $p = .69$; MVN package, Korkmaz et al., 2019). A regression slope test was used to verify the null hypothesis that the slope of the regression line was equal to zero.

In experiments 2 and 3, in-house R scripts were used to generate random and phonetically reduced lists for comparison purposes. In both experiments, ten lists were generated per condition. For the evaluation of the PPD index reliability in reduced lists, we combined several independent metrics to account for the specificity of our dataset. Since non-

canonical data structure can heavily impact the classifier's performance, several features related to the data were inspected prior to the analyses including class imbalance, degree of overlap between classes and class structure.

With a Shannon entropy index of 0.92 (0 being indicative of low entropy and 1 of high entropy, i.e., equal probabilities of occurrence for each class), the analyses indicated that the dataset was mildly imbalanced with a ratio of 1 to 3, that is, the event 'healthy' (the minority class) having approximately 33% chance of occurring. This suggests a possibility of a slight bias towards the majority class ('patient'). In terms of distance estimation between classes of data, the PPD score distribution analyses in the original dataset revealed an overlap index of 0.5 (where 0 indicates a perfect separability and 1 a complete overlap, Ridout and Linkie, 2009) implying that 50% of each class distribution is shared. Finally, there were differences in variance between 'heathy and 'patient' classes with a higher variance for 'patients', the majority group (σ = 0.38 vs. σ = 0.05). Now, unlike in most studies handling imbalanced data such as fraud or disease detection, where the minority class not only tends to be undersampled but also lacks a clear structure, variance differences identified in our dataset indicate that (i) the minority class ('healthy') is homogeneous and free of noise, and (ii) the majority class ('patient')—the actual target of the detection model—does exhibit greater dispersion around the mean, but is in turn sufficiently represented in the sample.

On the basis of the initial analysis, to assess the discrimination ability of the PPD score as a binary classifier (healthy vs. speech-impaired speaker) we first used the area under the receiver operating characteristics curve (AUC of ROC), the most popular discrimination metric for comparing the accuracy of independent clinical diagnostic tests. An ROC curve is obtained by plotting the proportion of true-positive rate (sensitivity, i.e., correct diagnosis) against false-positive rate (incorrect diagnosis) at each classification threshold. AUC summarizes performances across all thresholds and provides a scalar measure of estimated probability that a randomly selected speech-impaired speaker will be ranked as such above a randomly selected healthy person. An ideal test with AUC = 100% would have 100% true-positives with zero false-positives across all thresholds (top-left corner of the ROC curve). A test with poor discrimination ability will have AUC around 50%, that is, the classifier has classified at random. In the present study, AUC was calculated for original ROC curves and AUC statistics were obtained by means of functions available in the R pROC package (Robin et al., 2011).

While ROC curves and AUC statistics provide a useful summary of the classifier's performance, they are sensitive to imbalanced data (e.g., Saito & Rehmsmeier, 2015). For this reason, two alternative metrics were used to complete the analyses of classifier performance. The first of them, Balanced Accuracy, normalizes true positive (Sensitivity) and true negative predictions (Specificity) by the sum of positive and negative samples and computes the arithmetic mean of the two:

$$Balanced\ Accuracy = \frac{\text{Sensitivity} + \text{Specificity}}{2} \quad (1)$$

Balanced Accuracy score was computed using a function available in *yardstick* package in R (Kuhn & Vaughan, 2020). The second metric, the F1-score relies on the notions of precision (predicted positives divided by all positive predictions) and recall (the proportion of successfully identified positives) and combines them into one metric that corresponds to the weighted harmonic mean of the two:

$$F1\ score = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

Both metrics require determining a cut-off point to estimate true positives and negatives, which is usually established at 0.5. Instead of using the generic threshold, we opted for computing an optimal threshold value for each particular model using the Youden index formula (3). The Youden score provides the best trade-off between sensitivity and specificity and allows to account for differences attributable to pseudoword lists' specificity. Once established, the threshold was expressed in terms of PPD score below which speakers were considered as belonging to the 'healthy' group.

$$J = \text{Sensitivity} + \text{Specificity} - 1 \quad (3)$$

While Balanced Accuracy and F1-metric were used to assess the PPD classifier's performance, we also aimed at concurrent validation of the PPD index, that is, at establishing to which extend the PPD-based intelligibility index was related to an independent clinical metric of intelligibility, typically used in speech disorders as a gold standard, namely the Severity score (Balaguer et al., 2019). Severity scores ranging from 0 (severe disability) to 10 (normal speech) were given by six clinicians who listened to each patient reading a text or describing a picture, with the six scores for each item then averaged. The metric was available for a subset of speakers (n = 105). Following prior normality assessment, correlation analyses were performed to test the hypothesis that PPD and Severity scores will be highly correlated, independently of the sample size.

Experiment 1

The goal of experiment 1 was to exclude the possibility that variability in intelligibility estimates might be due to lack of consistency between the lists randomly generated from the pseudoword corpus rather than differences between

participant groups. Given the careful thought applied to selecting the phonetic criteria underlying the composition of the lists, we expected that PPD scores elicited based on two randomly generated lists would be very strongly correlated.

Results of the correlation analyses confirmed our prediction by showing a very strong positive relationship between the two lists as illustrated in Fig. 1. The fitted β_1 slope was 1.01 with an interval of (0.94, 1.08). The regression slope test showed that the slope was significantly different from 0 and thus confirmed that the linear relationship between PPD scores obtained from each pseudoword list was statistically significant (F(1-18) = 907.5, p = .000).

Experiment 2

An earlier study (Laaridh et al., 2018) based on the PPD speech corpus proposed a drastic reduction of the original 52 lists to subsets of ten pseudowords, representing 20% of the original elicitation materials and approximately 7 s of speech. The reduction was randomly performed, that is, without considering the phonetic content of the ten-item sublists. The sublists were then read aloud by healthy and head and neck cancer patients and the recordings were rated for intelligibility perceptually by human raters and automatically by means of software. Correlation between the two sets of intelligibility ratings was then checked by means of two measures of fit, the R coefficient and the root mean square error. The study revealed that the two ratings were significantly positively correlated for the ten-item sublists (R = 0.75 and RMSE = 2.719) but lost 0.09 in collinearity as compared to the full 52-item list

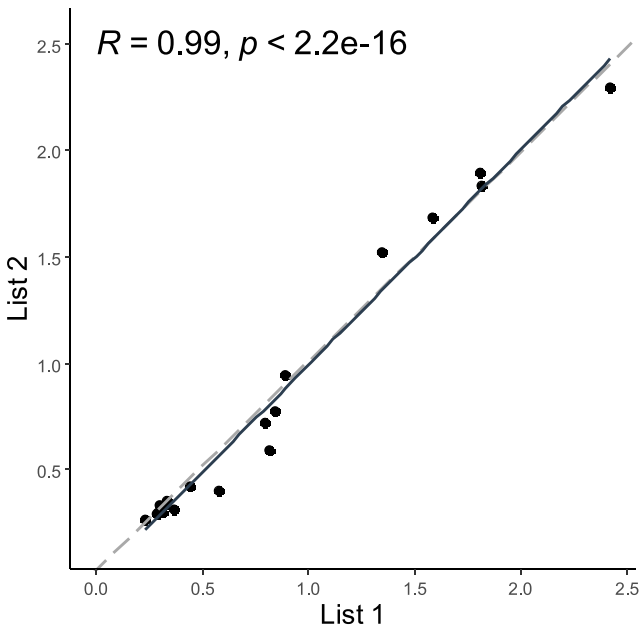

Fig. 1 Correlation between PPD scores elicited based on two randomly generated pseudoword lists. The *black line* represents the regression slope

(R = 0.84, RMSE = 2.339). Whereas—surprisingly—this result indicates that the overall predictive ability of the index was not highly sensitive to the lack of data, the outcome was instead heavily dependent on the list, which suggests that the acoustic and phonetic composition of pseudowords matters for intelligibility measures.

In the present study, we expanded on this initial evidence by systematically testing the relationship between a randomly reduced sample size and the accuracy of the instrument. While Laaridh et al. limited their analyses to sublists of ten items, we sought to examine how reliability would evolve as datasets were incrementally reduced. Accordingly, random sublists were generated from the original set of 52 in a stepwise manner going from 35 to five items per list. The robustness of the intelligibility index was examined in terms of two indicators: the AUC values for each reduced subset, and the strength of correlation between the full and reduced sets as indicated by Spearman coefficient. These analyses were completed by carrying out a stability assessment. For that purpose, rather than averaging across sublists, which would likely bias accuracy markers upward (see Laaridh et al., 2018: 2946), we chose to report ranges of values for both accuracy metrics. This is because, as argued in the introduction, optimality requires that all three criteria—accuracy, consistency, and stability—are fulfilled.

Our predictions for experiment 2 were threefold. First, if the PPD index preserves its strong ranking ability even when based on considerably reduced material, we would expect to see overall high positive collinearity between the full and reduced lists and overall high AUC values. Second, we would nonetheless expect these accuracy markers to improve with greater sample size. And thirdly we predicted that sample size would have considerable impact on the stability of the results. Specifically, smaller sample size should be associated with greater variability in the PPD scores' ability to discriminate as compared to larger samples.

The results summarized in Table 3 indicate that, overall, the PPD scores based on reduced vs. full lists were highly correlated with correlation coefficients ρ ranging from 0.85 to 0.98. Similarly, the observed AUC values going from 0.86 to 0.94 are indicative of a high to very high discriminating ability on the part of the PPD intelligibility index. Overall, the accuracy markers improve with increasing sample size. However, as expected, accuracy indicators plotted on Fig. 2 show greater instability of the results for smaller sample sizes (5–15) and suggest that accuracy results get more stable starting from around 40–50% of the original data (n = 20–25). Taken together, these results show that group classification computed from sublists representing less than 40% of the original elicitation materials are generally accurate but not optimal, due to the relative instability of the accuracy markers.

Experiment 3

As a follow-up to the previous results, experiment 3 aimed to explore whether data reduction based on phonetically motivated criteria would allow a more stable and reliable result when compared to the reference dataset and to an arbitrarily reduced list of the same size.

Our rationale for the optimal data reduction was based on combined statistical and phonetic considerations and guided by an underlying assumption that linguistic materials of greater complexity—in terms of both speech encoding and speech decoding—would make it possible to assess intelligibility more precisely. First, because the relative contribution of consonants and vowels to intelligibility perception is still a matter of vigorous debate (Kewley-Port et al., 2007; Nazzi & Cutler, 2019; Stilp & Kluender, 2010), we chose to act on consonants rather than on vowels for statistical reasons. Specifically, there is more uncertainty when choosing among 36 possible consonants (16 as singletons plus at least 18 in clusters) than among eight possible vowels, which we assumed would have an impact on listeners' decisions. Our second criterion was phonetically motivated. We felt that consonant clusters would be more relevant for intelligibility assessment than singletons because they are articulatorily and perceptually more complex. Cluster articulation involves rapid changes in vocal tract constrictions and are particularly challenging for speakers with dysarthric impairment (Kuruvilla-Dugdale et al., 2018). Articulatory complexity is in turn reflected in their acoustic signatures and therefore decoding complex syllables would also imply a higher processing cost for the listener than simple syllables. Our third criterion was that of representativity. To meet the representativity requirement, we wanted each consonant to have an equal chance of occurrence in the reduced sample. Because certain French consonants do not cooccur, we allowed singletons and clusters to appear in the C1 position but restricted the C2 position to clusters only. We preferred to perform reduction on the second rather than on the first syllable of the pseudoword because the intervocalic consonant cluster might create additional processing complexity for the listener by being assignable to either of the syllables. For example, a tautosyllabic cluster such as /d / always forms a syllable onset, while for a heterosyllabic cluster such as /kt/ the first consonant is assigned as a coda to the first syllable, and /t/ as onset to the second syllable.

By removing 36 singletons, we obtained phonetically reduced sublists per speaker containing 16 pseudowords of complex syllabic structure (henceforth, 16 Phon). For comparison, an alternative sublist containing 24 pseudowords characterized by a simple syllabic structure (henceforth, 24 CVCV) and ten randomly generated sublists of 16 items each (1–10 Rand) were saved for further analyses. We expected that from these cross-list comparisons the phonetically controlled materials would yield an optimal performance from the classifier. In addition to the accuracy criteria used in experiment 2, we also performed correlation analyses with Severity scores. The results reported below are organized according to the accuracy parameter tested.

List consistency assessment The scatterplots provided in Fig. 3 summarize our list consistency analyses, based on AUC statistics (see Table 4). Spearman's correlation analyses for not normally distributed data revealed that scores obtained from all reduced samples were overall very highly correlated

Table 3 Accuracy values for ten randomly generated sublists of varying size. In *bold*, the highest AUC values. For comparison, AUC value for the original dataset equals 94.08%

Accuracy indicator	*n*	Indicator values per sample size										Range
ρ	5	0.87	0.86	0.88	0.89	0.89	0.91	0.89	0.89	0.85	0.87	0.85–0.91
AUC		91%	86%	85%	91%	86%	88%	91%	88%	85%	87%	85–91%
ρ	10	0.91	0.92	0.94	0.94	0.93	0.92	0.91	0.94	0.92	0.93	0.91–0.94
AUC		92%	92%	92%	89%	90%	93%	91%	92%	88%	**94%**	88–94%
ρ	15	0.96	0.94	0.95	0.96	0.95	0.95	0.95	0.95	0.95	0.96	0.94–0.96
AUC		93%	89%	92%	92%	91%	91%	93%	90%	90%	91%	89–93%
ρ	20	0.96	0.97	0.96	0.95	0.97	0.96	0.97	0.97	0.97	0.96	0.95–0.97
AUC		91%	92%	92%	**94%**	92%	**94%**	**94%**	92%	93%	93%	91–94%
ρ	25	0.97	0.97	0.97	0.97	0.96	0.96	0.97	0.98	0.96	0.97	0.96–0.98
AUC		91%	93%	**94%**	92%	93%	**94%**	92%	**94%**	93%	92%	91–94%
ρ	30	0.97	0.98	0.97	0.98	0.98	0.97	0.98	0.97	0.97	0.98	0.97–0.98
AUC		93%	92%	93%	**94%**	93%	93%	93%	92%	92%	93%	92–94%
ρ	35	0.97	0.97	0.98	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.97–0.98
AUC		91%	**94%**	90%	90%	93%	93%	93%	93%	93%	92%	91–94%

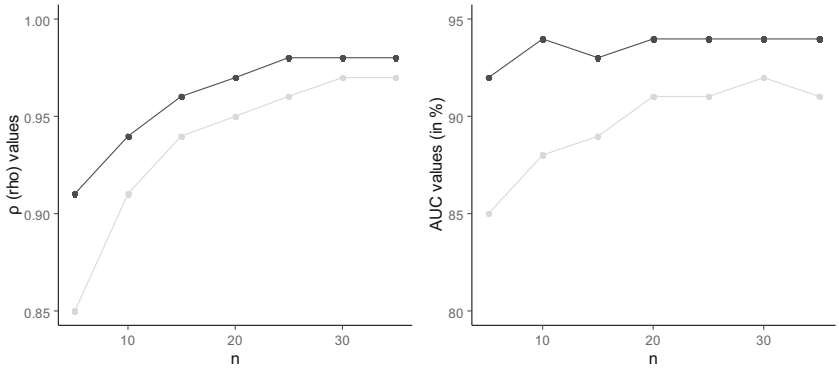

Fig. 2 Ranges of accuracy indicators depending on the sample size (n = 5–35). *Light grey* represents minimal values, *dark grey* represents maximal values for each n

with those generated from the original material[1], with the highest collinearity coefficient for the phonetically reduced simple sublist (ρ = 0.98, top right panel) and the lowest coefficient for the worst of the randomly generated sublists (ρ = 0.96, bottom right panel). It should be noted, however, that the best sublist represents 46% of the original materials as opposed to all the other compared sublists (each one representing 31% of the original dataset). Overall, the fact that the intelligibility scores obtained from reduced lists (n = 16) so closely mirror those from the original dataset (n = 52) provide a strong argument for the reliability of the PPD index even when generated from a reduced dataset.

Classifier performance assessment Table 4 provides a summary of classifier accuracy for all lists studied based on three indicators: AUC, Balanced Accuracy scores and F-scores. We observe that the models generated on the phonetically complex sublist and the original dataset are equivalent in terms of area under the curve, above 94% for both (z = 0.10, p = .920). This result indicates that the ranking ability of the PPD intelligibility score is as reliable when performed on a phonetically reduced sample size as it is when based on the original larger sample. The sublist that retained simple syllables, which was highly correlated with the original dataset (see above), is significantly less accurate than the original set (z = 2.96, p = .003). Balanced Accuracy and F indexes for skewed class distributions yield similar results: the phonetically reduced dataset exhibit the second highest score and systematically outperforms the simple syllable dataset (see Table 4).

Turning to randomly reduced pseudoword lists, the AUC values for ten lists of the same length range from 93.69 to 88.75% with a mean AUC of 91.57%. The dispersion of these results, as well as higher standard errors and wider confidence intervals, indicates relative instability in the classifying performance when the linguistic materials are arbitrarily reduced. When compared with the discriminatory power of the phonetically reduced complex and original models, the least reliable of the random models tested (model 7, see Table 4) is significantly less discriminating than either of them (z = 2.6098, p = .009 and z = 3.4417, p = .001, respectively). Balanced Accuracy index (ranging from 0.8476 to 0.8844) and F-score (from 0.8477 to 0.9240) show similar instability of the results. Taken together, this outcome reveals a risk related to the classifying ability of an intelligibility score based on arbitrarily generated lists. That is, if discrimination is based on randomly extracted pseudowords, it may or may not result in a comparably reliable classifier. ROC curves illustrating these discrepancies are given in Fig. 4.

Consistency with Severity index To complete the analyses reported above, Spearman's correlation coefficients were computed to assess the strength of the relationship between the PPD speech intelligibility and Severity scores (i.e., a subjective assessment by clinicians), depending on the linguistic material. As illustrated in Fig. 5, moderately strong negative correlations were observed between severity measures and all the intelligibility indexes, implying that loss in intelligibility (associated with a higher PPD score) is significantly correlated with increase in severity (a low severity index). The intelligibility index obtained from the phonetically complex sublist was the most strongly correlated with the severity index (ρ = – 84, n = 16, p = .000).

Discussion and conclusion

The analyses reported in this paper confirmed our initial hypothesis that phonetically motivated data reduction would make it possible to optimize the sample size used in intelligibility assessment. We have shown by means of independent accuracy markers—collinearity with the original dataset, collinearity with an alternative speech impairment index and accuracy analyses—that reducing the original data by 30% on the basis of phonetic criteria related to phonotactic complexity

[1] Correlation coefficients were slightly lower for best and worst lists as selected in terms of Balanced Accuracy and F-scores. Accordingly, the worst list (List 1), ρ = 0.95, and the three top lists (Lists 3, 5, and 6) ρ = 0.95, ρ = 0.96 and ρ = 0.94, respectively.

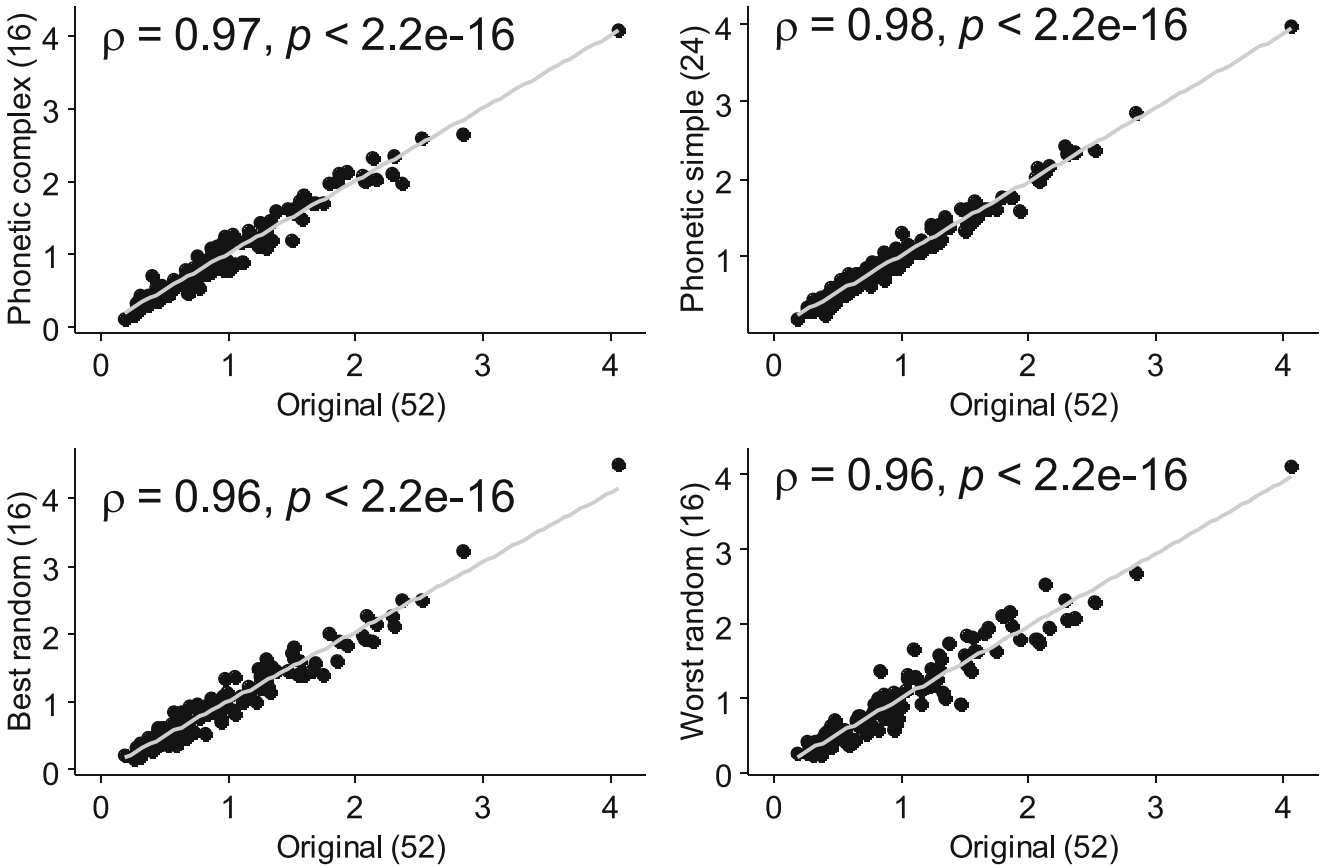

Fig. 3 Correlations between the original 52-item and reduced datasets and respective collinearity coefficients and significance

will provide an optimal intelligibility metrics to discriminate between speech disordered populations (here, related to speech sequelae of HNC) and healthy speakers. Such phonetically motivated data reduction proved more reliable than arbitrary list reduction. This result has important clinical implications as it minimizes the time required to gather the speech sample while ensuring a statistically robust and stable result.

Table 4 Summary report of the model assessment with number of pseudowords in the dataset (*n*), area under the curve (AUC), related confidence interval (CI) and standard error (SE), Balanced Accuracy and F-scores

List	*n*	AUC of ROC			Balanced Accuracy score	F-score
		AUC in %	95% CI (DeLong)	SE		
Original	52	94.08%	0.8983–0.9857	0.00050	0.8841	0.9277
Reduced Phonetic Complex	16	94.19%	0.9060–0.9822	0.00038	0.8719	0.9146
Reduced Phonetic Simple	24	91.35%	0.8606–0.9664	0.00072	0.8476	0.8875
Reduced Random 1	16	91.51%	0.8622–0.9680	0.00073	0.8476	0.8477
Reduced Random 2	16	93.69%	0.8931–0.9806	0.00050	0.8536	0.8944
Reduced Random 3	16	92.77%	0.8796–0.9758	0.00060	0.8841	0.9125
Reduced Random 4	16	91.99%	0.8687–0.9712	0.00068	0.8720	0.8609
Reduced Random 5	16	90.77%	0.8511–0.9644	0.00084	0.8597	0.9240
Reduced Random 6	16	93.07%	0.8874–0.9740	0.00049	0.8841	0.9125
Reduced Random 7	16	88.75%	0.8240–0.9510	0.00105	0.8536	0.8944
Reduced Random 8	16	90.10%	0.8460–0.9560	0.00079	0.8536	0.9024
Reduced Random 9	16	92.48%	0.8769–0.9727	0.00060	0.8658	0.8917
Reduced Random 10	16	90.62%	0.8482–0.9642	0.00088	0.8597	0.8846

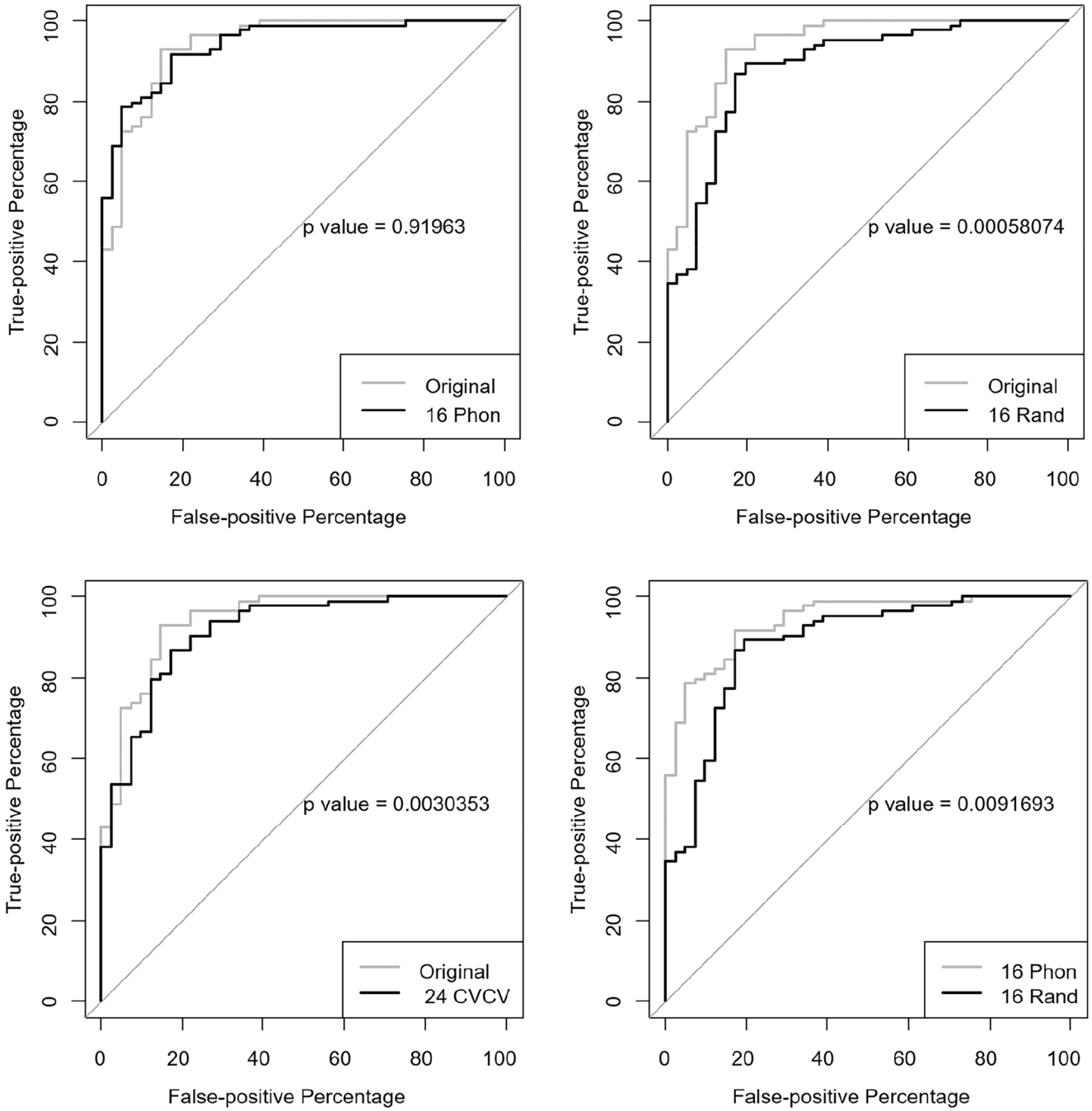

Fig. 4 ROC curve comparison for PPD speech intelligibility index. *Light grey line* represents a chance level (AUC = 0.5)

Our findings also have direct implications for speech intelligibility studies in disordered populations. They are in line with previous research that argued for phonetically structured materials for feature-based intelligibility assessment (see Introduction). Here, we add to the arguments previously put forward—such as the explanatory power of assessment based on such materials—that controlling for phonetic factors allows an optimal data reduction by selecting the linguistic units that seem to be the most relevant for revealing articulatory distortions characteristic of dysarthria. Our results, indicating that phonotactic complexity linked to within- and across-syllable consonant groups is a relevant unit to distinguish between speech-impaired and healthy speakers, are congruent with phonetic descriptions of dysarthric speech (Kent & Kim, 2011; Kim et al., 2010; Reilly & Spencer, 2013). Indeed, when the linguistic materials were limited to pseudowords of simple CVCV syllable structure, both the classifier performance and collinearity assessment yielded lesser accuracy.

However, the fact that some very short, randomly selected sublists consisting of ten pseudowords can exhibit

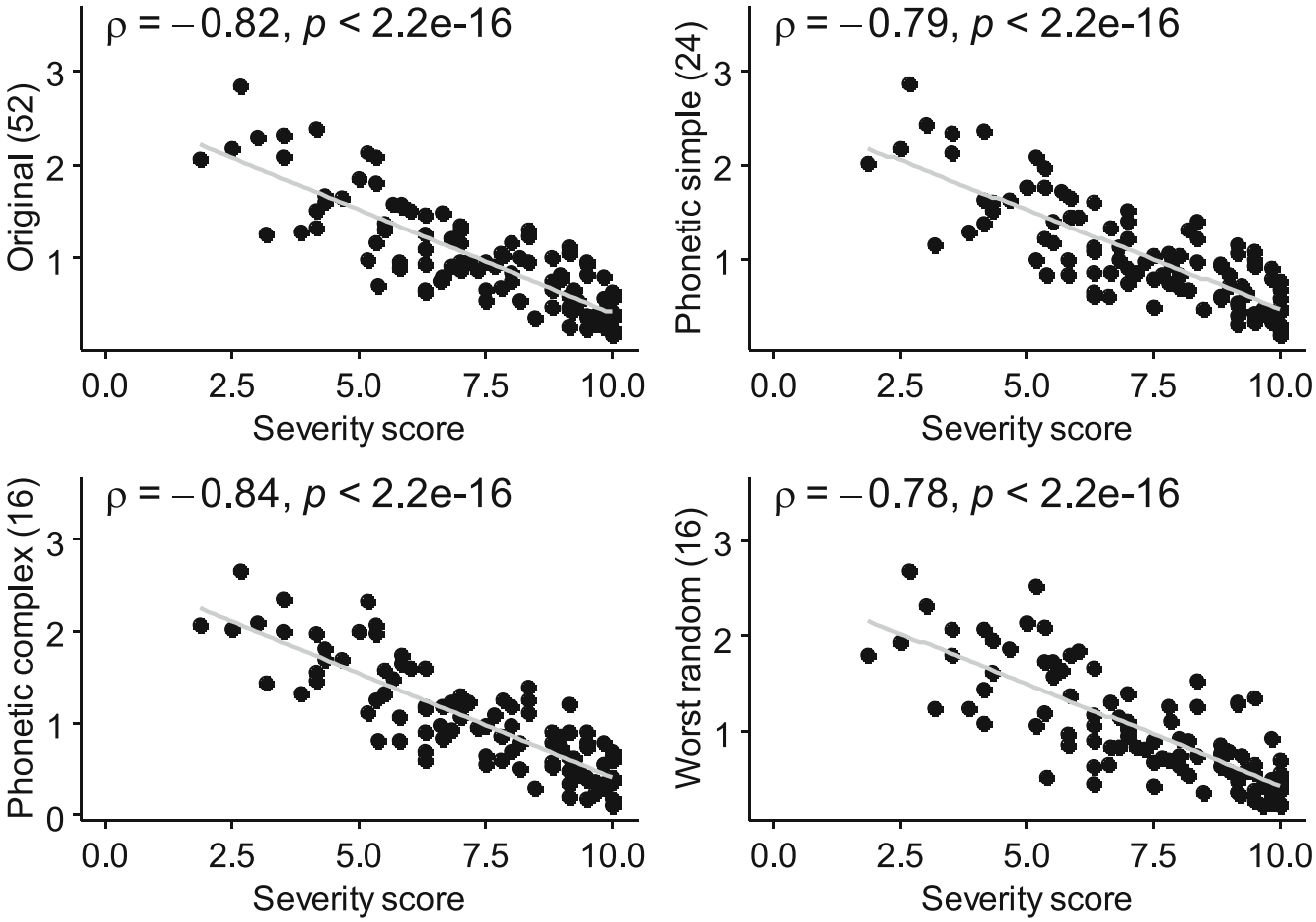

Fig. 5 Linear relationship between the PPD and Severity scores for the original 52-item and reduced sublists with their respective correlation coefficients and significance

comparably high accuracy (see Table 3) is indicative that there may be alternative phonetic criteria for list composition that would provide equally reliable intelligibility estimates. Therefore, alternative phonetically motivated hypotheses for data reduction could be considered, and future work should examine other phonetically and psycholinguistically relevant variables that are likely to reflect intelligibility loss such as vowel characteristics or frequency patterns. To this end, error analyses in sample transcription might provide insights about the hierarchy of processes involved in acoustic phonetic decoding. In addition, future research should test the robustness of the PPD classifier on phonetically reduced lists using automatic analytical tools such as those within the i-vector paradigm and support vector regression-based models.

Open Practices Statement

Neither of the studies reported in this article was formally preregistered. The data have not been made available on a permanent third-party archive because our Institutional Review Board ruled that we could not post the data; requests for the data can be sent via e-mail to the lead author.

Acknowledgements This work was supported by Grant n°2014-135 from Institut National pour le Cancer (INCA) lead by Pr Virginie Woisard at University Hospital of Toulouse and by 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 led by Jérôme Farinas at the IRIT, Toulouse, France.

Availability data The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Allen, K., Alais, D., & Carlile, S. (2012). A collection of pseudo-words to study multi-talker speech intelligibility without shifts of spatial attention. *Frontiers in Psychology*, *3*(MAR), 15–17. https://doi.org/10.3389/fpsyg.2012.00049

André, C., Ghio, A., Cavé, C., & Teston, B. (2003). PERCEVAL: A Computer-Driven System for Experimentation on Auditory and Visual Perception. *Proceedings of International Congress of*

Phonetic Sciences (ICPhS), *1*, 1421–1424. http://arxiv.org/abs/0705.4415

Astésano, C., Balaguer, M., Farinas, J., Fredouille, C., Gaillard, P., Ghio, A., … Woisard, V. (2018). Carcinologic speech severity index project: A database of speech disorder productions to assess quality of life related to speech after cancer. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 4265–4271.

Auzou, P., & Rolland-Monnoury, V. (2006). *BECD : Batterie d'Evaluation Clinique de la Dysarthrie*. Isbergues, France: Ortho Edition.

Balaguer, M., Boisguérin, A., Galtier, A., Gaillard, N., Puech, M., & Woisard, V. (2019). Assessment of impairment of intelligibility and of speech signal after oral cavity and oropharynx cancer. *European Annals of Otorhinolaryngology, Head and Neck Diseases*, *136*(5), 347–352.

Barreto, S., Dos S., & Ortiz, K. Z. (2010). Intelligibility: effects of transcription analysis and speech stimulus. *Pro-Fono Revista de Atualização Científica*, *22*(2), 125–130. https://doi.org/10.1590/s0104-56872010000200010

Bechet, F. (2001). LIA_PHON: un système complet de phonétisation de textes. *Traitement Automatique Des Langues - TAL*, *42*(1), 47–67.

Enderby, P. (1983). *Frenchay dysarthria assessment*. College Hill Press.

Ghio, A., Lalain, M., Giusti, L., & Woisard, V. (2020). How to compare automatically two phonological strings: application to intelligibility measurement in the case of atypical speech. *LREC Language Resource and Evaluation Conference, Marseille*, 1689–1694.

Haley, K. L., Jacks, A., Richardson, J. D., & Wambaugh, J. L. (2017). Perceptually salient sound distortions and apraxia of speech: A performance continuum. *American Journal of Speech-Language Pathology*, *26*(2S), 631–640. https://doi.org/10.1044/2017_AJSLP-16-0103

Kent, R. D., & Kim, Y. J. (2011). The assessment of intelligibility in motor speech disorders. In A. Lowit & R. D. Kent (Eds.), *Assessment of motor speech disorders* (pp. 21–37). San Diego, CA: Plural Publishing.

Kent, R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, *54*(4), 482–499. https://doi.org/10.1044/jshd.5404.482

Kewley-Port, D., Burkle, T. Z., & Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *122*(4), 2365–2375. https://doi.org/10.1121/1.2773986

Kim, H., Martin, K., Hasegawa-Johnson, M., & Perlman, A. (2010). Frequency of consonant articulation errors in dysarthric speech. *Clinical Linguistics and Phonetics*, *24*(10), 759–770.

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2019). *MVN: An R Package for Assessing Multivariate Normality*. https://journal.r-project.org/archive/2014-2/korkmaz-goksuluk-zararsiz.pdf

Kuhn, M., & Vaughan, D. (2020). *Yardstick: Tidy Characterizations of Model Performance*. https://cran.r-project.org/package=yardstick

Kuruvilla-Dugdale, M., Custer, C., Heidrick, L., Barohn, R., & Govindarajan, R. (2018). A phonetic complexity-based approach for intelligibility and articulatory precision testing: A preliminary study on talkers with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, *61*(9), 2205–2214. https://doi.org/10.1044/2018_JSLHR-S-17-0462

Laaridh, I., Fredouille, C., Ghio, A., Lalain, M., & Woisard, V. (2018). Automatic evaluation of speech intelligibility based on i-vectors in the context of head and neck cancers. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2943–2947. https://doi.org/10.21437/Interspeech.2018-1266

Lagerberg, T. B., Johnels, J. Å., Hartelius, L., & Persson, C. (2015). Effect of the number of presentations on listener transcriptions and reliability in the assessment of speech intelligibility in children. *International Journal of Language and Communication Disorders*, *50*(4), 476–487. https://doi.org/10.1111/1460-6984.12149

Lalain, M., Ghio, A., Giusti, L., Robert, D., Fredouille, C., & Woisard, V. (2020). Design and development of a speech intelligibility test based on pseudowords in French: Why and how? *Journal of Speech, Language, and Hearing Research*, *63*(7), 2070–2083.

Leplège, A., Ecosse, E., Pouchot, J., Coste, J., & Perneger, T. (2001). *Le questionnaire MOS SF-36 : manuel de l'utilisateur et guide d'interprétation des scores*. Paris: Estem.

Miller, N. (2013). Measuring up to speech intelligibility. *International Journal of Language and Communication Disorders*, *48*(6), 601–612. https://doi.org/10.1111/1460-6984.12061

Nazzi, T., & Cutler, A. (2019). How Consonants and Vowels Shape Spoken-Language Recognition. *Annual Review of Linguistics*, *5*(1), 25–47. https://doi.org/10.1146/annurev-linguistics-011718-011919

R Core Development Team. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/

Rebourg, M., Lalain, M., Ghio, A., Fredouille, C., Fakhry, N., & Woisard, V. (2020). Évaluer l'intelligibilité, mots ou pseudo -mots ? Comparaison entre deux groupes d'auditeurs. *Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1:* Journées d'Études sur la Parole, 2020, Nancy, France (pp. 543–551).

Reilly, K. J., & Spencer, K. A. (2013). Sequence Complexity Effects on Speech Production in Healthy Speakers and Speakers with Hypokinetic or Ataxic Dysarthria. *PLoS ONE*, *8*(10), 1–14. https://doi.org/10.1371/journal.pone.0077450

Ridout, M.S., Linkie, M. Estimating overlap of daily activity patterns from camera trap data. *Journal of Agricultural, Biological, and Environmental Statistics 14*, 322–337 (2009). https://doi.org/10.1198/jabes.2009.08038

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 1–17. https://doi.org/10.1186/1471-2105-12-77

RStudio Team. (2015). *RStudio: Integrated Development Environment for R*. http://www.rstudio.com

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, *10*(3), 1–21. https://doi.org/10.1371/journal.pone.0118432

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*, *110*(4), 474–494. https://doi.org/10.1037/0096-3445.110.4.474

Stilp, C. E., & Kluender, K. R. (2010). Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(27), 12387–12392. https://doi.org/10.1073/pnas.0913625107