Have a Cake and Eat it Too: Assessing Discrimination Performance of an Intelligibility Index Obtained from a Reduced Sample Size

Anna K Marczyk, A Ghio, Muriel Lalain, Marie Rebourg, C. Fredouille, Virginie Woisard

Have a Cake and Eat it Too: Assessing Discrimination Performance of an Intelligibility Index Obtained from a Reduced Sample Size

A. Marczyk[1], A. Ghio[1], M. Lalain[1], M. Rebourg[1], C. Fredouille[2], V. Woisard[3, 4]
[1]Aix-Marseille Université CNRS-LPL, Aix-en-Provence, France; [2] LIA, Avignon University, France; [3]Service ORL, CHU Larrey; [4]URI Octogone-Lordat, Toulouse, France
{anna.marczyk. alain.ghio. muriel.lalain. marie.rebourg}@lpl-aix.fr, corinne.fredouille@univ-avignon.fr, woisard.v@chu-toulouse.fr

Abstract
This paper investigates random vs. phonetically motivated reduction of linguistic material used in an intelligibility task in speech disordered populations and the subsequent impact on the discrimination classifier quantified by the area under the receiver operating characteristics curve (AUC of ROC). The comparison of obtained accuracy indexes shows that when the sample size is reduced based on a phonetic criterium—here, related to phonotactic complexity—, the classifier has a higher ranking ability than when the linguistic material is arbitrarily reduced. Crucially, downsizing the linguistic sample to about 30% of the original dataset does not diminish the discriminatory performance of the classifier. This result is of significant interest to both clinicians and patients as it validates a tool that is both reliable and efficient.

Keywords: speech intelligibility index, discrimination ability, speech disorders, ROC curve

1. Introduction

Speech intelligibility measure[1] represents a scalar metric that reflects how well a message is decoded by a listener and gives a reliable estimate of word-form impairments, a cardinal symptom of common speech disorders such a dysarthria. It serves multiple purposes in clinical settings, from being used as a diagnostic tool for patient discrimination and for determining the degree of impairment, to serving as an intervention monitoring indicator. It has been typically obtained based on linguistically controlled single word and sentence material read by patients and orthographically transcribed by independent listeners (Kent, Weismer, Kent, & Rosenbek, 1989). Since interferences due to word familiarity are an important caveat of using word material for clinical speech assessment, intelligibility metrics have been more recently obtained from pseudoword lists (e.g. Allen, Alais, & Carlile, 2012). A phonetically controlled pseudoword material designed for French-speaking speech disordered population has been recently proposed for the acoustic-phonetic decoding task with the aim of circumventing confounds due to the lexical bias, such as phonemic restauration or learning effects (Perceived Phonological Deviation score, henceforth PPD, Ghio et al., 2016, 2018; Lalain et al., under review).

Next to the type of linguistic material used for intelligibility assessment is the issue related to the trade-off between the index ranking ability and the volume of data needed to generate statistically reliable results. Clinical practice imposes considerable time constraints on practitioners and obtaining a reliable and efficient speech performance classifier becomes critical. While PPD intelligibility score proved to be highly performant in discriminating between healthy and carcinologic speakers (Ghio et al., 2018), it was generated from lists of 52 pseudowords, a sample that requires an important amount of time from a clinician's perspective. For comparison, the BECD dysarthria assessment battery (Auzou & Rolland-Monnoury, 2006), widely used by French-speaking speech therapy practitioners, generates intelligibility scores from 10 words and 10 sentences, randomly selected from lists of 50 items. Reducing testing time is equally important for patients. Patients' fatigue is a common reason to leave the task incomplete. Thus, the general objective of this paper is to examine the impact of reduced sample material on PPD index predictive ability in order to find an optimal tradeoff between the classifier performance and efficiency. In the remaining sections of the introduction we will briefly present the specifics of the PPD intelligibility measure and discuss the data reduction criteria.

1.1 PPD Speech Intelligibility Score

Below we summarize the aspects that are necessary to contextualize the present goals, the reader will find a detailed description of PPD development in relevant bibliography. The linguistic material proposed consists of 52 bisyllabic 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 either to 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 to generate 90 000 pseudowords (after exclusion of semantically meaningful items), a database from which equivalent pseudoword lists are generated. Table 1 provides a summary of the consonants and vowels retained for the corpus. The number of pseudowords on the final list (n=52) aims at ensuring high robustness of the proposed metric and is phonetically motivated. The material is robust in that it allows obtaining multiple samples of each speech sound so that each list is equally representative of the French sound system. Specifically, each consonant appears at least twice in each position, as

[1] We use the term speech intelligibility measure in a broad sense to refer to the listener's message decoding ability which is not to be confused with the Speech Intelligibility Index (SII), a physical measure of speech intelligibility based on acoustic properties of the speech signal.

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020) , pages 1784–1788
Marseille, 11–16 May 2020

singleton and in a consonant cluster, while each vowel appears at least six times in each syllable.

Position	Segmental content
C_1=singleton	p t k b d g v z ʒ f s ʃ ʁ l m n ŋ j
C_1=cluster	pʁ tʁ kʁ gʁ bʁ fʁ pl kl fl st bl sk sp gl dʁ ps
V_1	a i y u O E ɑ̃ ɛ̃
C_2=singleton	p t k b d g v z ʒ f s ʃ ʁ l m n ŋ j
C_2=cluster	st ks ʁd ʁs kt ʁn pl gʁ dʁ kl ʁj lt ʁv vʁ gz ʁp tʁ ʁt bl ʁm pʁ kʁ sk bʁ sp ʁk fʁ fl ʁb gl ps pt
V_2	a i y u O E ɑ̃ ɛ̃

Table 1. Summary of vowels and consonants that can appear in each of the phonetic contexts. Capital letters represent archiphonemes, that is, a class of phonemes sharing all but one feature (here, vowel height).

Because several singletons (such as /ʒ/ or /ʃ/) do not form frequent clusters with another consonant, single consonants are set to come out twice in C_1 and C_2 positions. Figure 1 depicts consonant distribution within the general pseudoword structure. Table 2 shows an example of a pseudoword list.

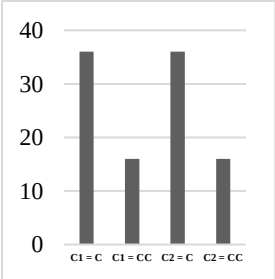

Figure 1. Distribution of singletons and consonant clusters on each PPD list. Y-axis represents the number of occurrences.

crampant	fevo	quinfant	flaspou	
plouniant	troucha	suptu	vabla	baillu
ratri	rougli	nougu	touflant	griti
inrtin	dibro	yango	zucrou	quebo
gavi	pufru	scuchu	psoussa	chanjin
bijo	blouillu	fampsi	madin	lupou
tanli	niascu	pimprant	climbou	
storquin	brori	chansin	jindou	
spucou	glima	prinrmo	gomou	droto
yezant	vefin	zelin	jezant	dorba
nioniou	mera	frina	lina	siqui

Table 2. Example of a pseudoword list in orthography.

The corpus designed in this way was used in the framework of CS2I project (Carcinologic Speech Severity Index, Astésano et al., 2018), within which 126 speech samples were obtained from 85 carcinologic patients and 41 healthy speakers. Each sample (52 pseudowords) was transcribed orthographically by three independent naive listeners. The task is termed acoustic-phonetic decoding[2]. Intelligibility scores were computed using the Levenshtein distance algorithm by comparing the expected and actual transcriptions. They correspond to the degree of dissimilarity (i.e. deviation), calculated in terms of distinctive features (maximal 6), between 35 French phonemes retained for the protocol. For example, for a PPD score equal to 3, the transcribed segment differs from the expected segment by three phonological features, that is, the higher the score value, the greater the distance between the expected and actual transcriptions and thus, the greater the intelligibility loss. The final PPD score was computed for each speaker by averaging scores across pseudowords and transcriptions.

1.2 List Reduction

For the reasons mentioned in the introduction, list reduction has become an urgent necessity. An earlier study (Laaridh, Fredouille, Ghio, Lalain, & Woisard, 2018) based on the PPD speech corpus proposed a drastic reduction of lists to subsets of 10 pseudowords, representing 20% of the original material, on which an automatic intelligibility prediction was carried out. The reduction was randomly performed, that is, without considering the phonetic content. The study revealed that while—surprisingly—the overall index predictive ability was not highly sensitive to the lack of data, the outcome was instead heavily dependent on the list, which suggests that acoustic and phonetic composition of pseudowords matters for intelligibility measures.

As a follow up to this work, in this paper we explore whether data reduction based on a phonetically motivated criterium allows 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 phonetically based reduction is as follows. First, we chose to act on consonants rather than on vowels because there is more uncertainty when choosing among 36 possible consonants (16 as singletons + min. 18 in clusters) than among 8 possible vowels, which we assumed would have an impact on listeners' decisions. Second, we considered that consonant clusters, whose production involves rapid changes in vocal tract constrictions with corresponding acoustic signatures, would have a higher processing cost for the listener than single consonants and therefore would be more relevant for the intelligibility assessment. In addition, in order to ensure the metric's reliability, we wanted the lists to meet the representativity requirement (see § 1.1), that is, ensure that each consonant is represented in the reduced sample. Because certain consonants do not cooccur, we accepted singletons and cluster to appear in the C_1 position but restricted the C_2 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 by being assigned to either of the syllables. For example, a tautosyllabic cluster such as /dʁ/ always forms

[2] Previous papers refer to PPD score by the name of the task used to elicit it (DAP score, from French *Décodage acoustico-phonétique*). The term PPD has replaced it and appears consistently in all later publications.

a syllable onset, while for an 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 (first column, Fig. 1), we obtained a phonetically reduced lists containing 16 pseudowords. Since the classifier seems robust even to important data reduction, we expect that a phonetically reduced list will produce a score of comparable predictive ability as for the original dataset. We hypothesize that a phonetically reduced list will prove more robust that a randomly reduced pseudoword list of similar length.

2. Method

PPD speech intelligibility score was computed for three different groups of datasets under comparison:

- The original dataset (n=52)
- A phonetically reduced dataset (n=16)
- 10 randomly reduced datasets (n=16 each)

To see whether the different lists were equivalent in their ability to separate the groups, we first carried out correlation analyses.

Turning to indices of accuracy, we used two independent measures to assess the discrimination ability of PPD as a binary classifier (speech disordered vs healthy group). First of them, the area under the receiver operating characteristics curve (AUC of ROC) is the most popular discrimination metric for comparing the accuracy of independent clinical diagnostic tests. A ROC curve is obtained by plotting the proportion of true-positive rate (correct diagnosis) against false-positive rate (incorrect diagnosis) at each classification threshold. AUC summarizes the performances across all thresholds and provides a scalar measure of estimated probability that a randomly selected patient will be ranked as patient 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 will do random ranking. ROC curves and AUC statistics were obtained by means of functions available in R package *pROC* (Robin et al., 2011).

To complete AUC analyses, we performed correlation analyses with an independent clinical measure, namely the Severity Score (Balaguer et al., 2019), available for a subset of speakers (n=105). Severity scores range from 0 (severe disability) to 10 (normal speech) and is based on a subjective assessment made by 6 clinicians who listen to the patient reading a text or describing a picture and propose a score on the 10 points scale. We expect both metrics to be highly correlated, independently of the sample size.

3. Results

2.1 Assessing Correlation Strength of the Classifier between Original and Reduced Samples

Pearson's correlation analyses revealed that scores obtained from reduced samples were overall very highly correlated with those generated from the original material, with the relationship being stronger between the phonetically reduced and the original models. Scatterplots provided in Figure 2 summarize these results. The fact that the intelligibility scores obtained from reduced lists (n=16) mirror those from the original dataset (n=52) provide a strong argument for the reliability of our procedure.

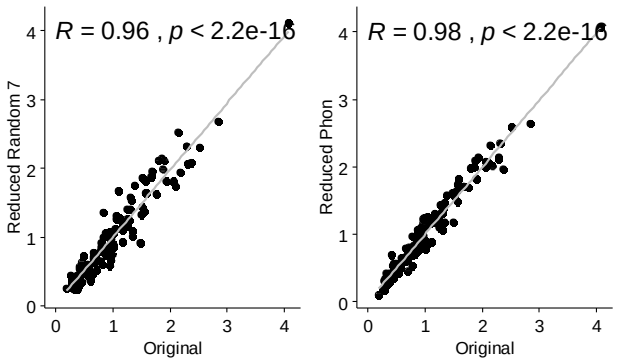

Figure 2. Linear relationship between the PPD intelligibility scores obtained for the weakest of randomly reduced datasets (model 7, left panel) and the phonetically reduced dataset (right panel) with their respective correlation coefficients and significance.

2.2 Assessing Classifier Performance with the Area under the ROC Curve Measure

We observe that the models generated on the phonetically reduced list and the original dataset are equivalent in terms of area under the curve, above 94% for both (z=-0.101, p=.009, see Table 3 and Figure 3, top panel). This result indicates that the ranking ability of the PPD intelligibility score is as reliable when performed on a substantially reduced sample size, as it is when based on the original larger material.

List	N	AUC in %	95% CI (DeLong)	SE
Original	52	94.20%	0.8983 - 0.9857	0.00050
Reduced Phonetic	16	94.41%	0.9060 - 0.9822	0.00038
Reduced Random 1	16	91.51%	0.8622 - 0.9680	0.00073
Reduced Random 2	16	93.69%	0.8931 - 0.9806	0.00050
Reduced Random 3	16	92.77%	0.8796 - 0.9758	0.00060
Reduced Random 4	16	91.99%	0.8687 - 0.9712	0.00068
Reduced Random 5	16	90.77%	0.8511 - 0.9644	0.00084
Reduced Random 6	16	93.07%	0.8874 - 0.9740	0.00049
Reduced Random 7	16	88.75%	0.8240 - 0.9510	0.00105
Reduced Random 8	16	90.10%	0.8460 - 0.9560	0.00079
Reduced Random 9	16	92.48%	0.8769 - 0.9727	0.00060
Reduced Random 10	16	90.62%	0.8482 - 0.9642	0.00088

Table 3. Summary report of the model assessment with number of pseudowords in the dataset (N), area under the curve (AUC) and related confidence interval (CI) and standard error (SE)

Turning to randomly reduced pseudoword lists, the AUC values for 10 lists of the same length ranges from 93.69% to 88.75% with a mean AUC of 91.57%. The dispersion of these results, as well as slightly higher standard errors, and wider confidence intervals indicate relative instability in the diagnostic performance when the linguistic material is arbitrarily reduced. When compared with the discriminatory power of the phonetically reduced and original models, the less reliable of the tested random models (model 7, see Table 3) is significantly less discriminant than either of them (z=2.6098, p=.009 and z=3.4417, p=.001 respectively, see Figure 3, middle and bottom panels).

This outcome reveals a risk related to the ranking ability of an intelligibility score based on arbitrarily generated lists. That it, if discrimination is based on randomly extracted pseudowords, it might, or it might not result in a comparably reliable classifier.

2.3 Assessing Correlation Strength between the Classifier and Severity Index

To complete the analyses reported above, Pearson's correlation coefficients were computed to assess the strength of the relationship between the speech intelligibility and severity scores, depending on the linguistic material. Moderately strong negative correlations were observed between severity measures and all the intelligibility indexes (see Figure 3) implying that loss in intelligibility (high PPD score) is strongly correlated with increase in severity (low severity index). The intelligibility index obtained from the phonetically reduced list was the most strongly correlated with severity index (r=-85, n=16, p=.000).

4. Conclusion

The most important result of the analyses reported above is that the speech intelligibility index obtained in the Acoustic Phonetic Decoding task is a reliable tool to discriminate between speech disordered populations (here, related to speech sequelae of head and neck cancers) and healthy speakers even if the sample size is reduced to about 30% of the original material, provided that the lists are not reduced arbitrarily but based on phonetic criteria of phonotactic complexity. 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. Future work may 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 hierarchy of processes involved in acoustic phonetic decoding. In addition, further work involves testing the robustness of the PPD classifier on phonetically reduced lists using automatic analyses, such as those within the i-vector paradigm and Support Vector Regression-based models.

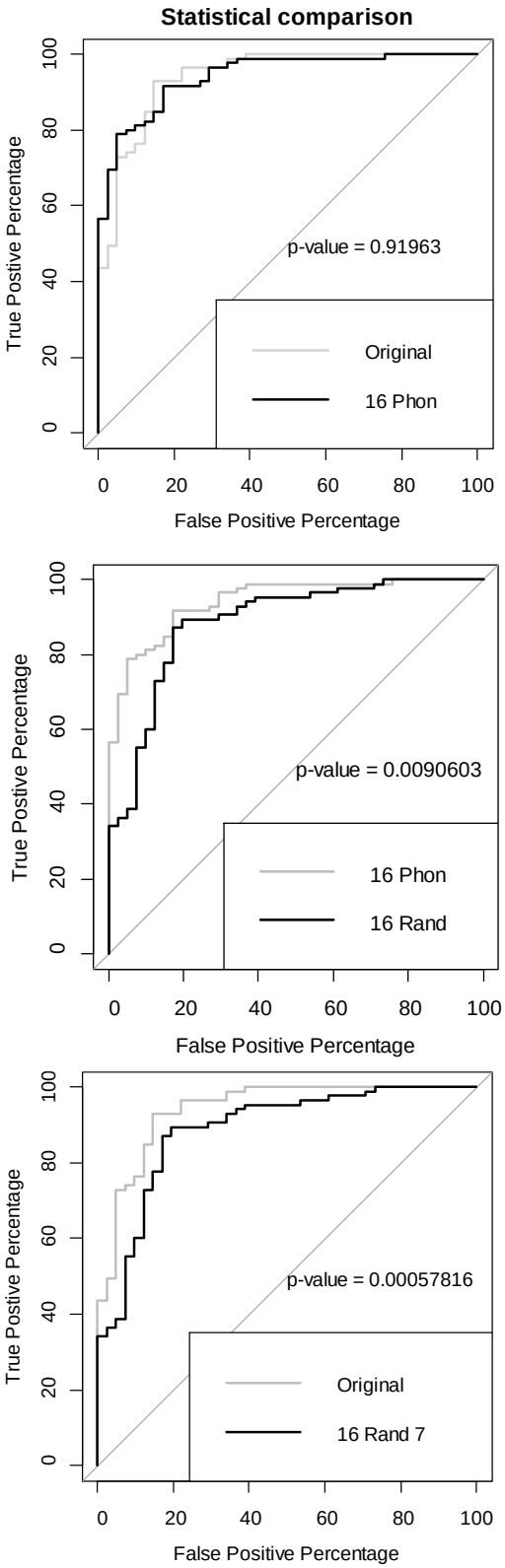

Figure 3. Statistical significance of ROC curve comparison for PPD speech intelligibility score. Top panel: original vs. phonetically reduced dataset; Middle panel: phonetically vs. randomly reduced datasets (model 7) of the same length; Bottom panel: original vs. the worst of randomly reduced datasets (model 7). Light grey line represents a chance level (AUC=0.5).

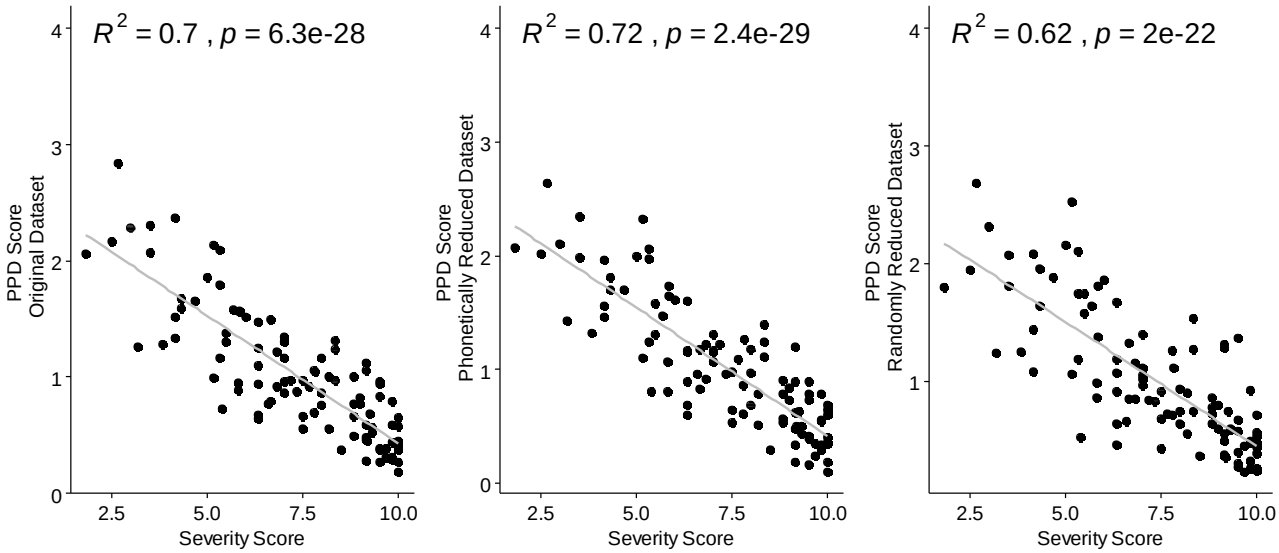

Figure 4. Linear relationship between the PPD score and Severity Score for the original (left panel), phonetically reduced (middle) and randomly reduced (right) datasets with their respective correlation coefficients and significance

5. 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. We thank Yussef Hmamouche for his help in R scripting.

6. Bibliographical 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

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.

Ghio, A., Giusti, L., Blanc, E., Pinto, S., Lalain, M., Robert, D., … Woisard, V. (2016). Quels tests d' intelligibilité pour évaluer les troubles de production de la parole ? *Journée d'étude Sur La Parole*, *4*(1), 589–596.

Ghio, A., Lalain, M., Giusti, L., Pouchoulin, G., Rebourg, M., Fredouille, C., … Robert, D. (2018). Une mesure d ' intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique To cite this version : HAL Id : hal-01770161 Une mesure d ' intelligibilité par décodage acoustico -phonétique de pseudo-mots dans le cas de par. *XXXIIe Journées d'Etudes Sur La Parole, LPL, 2018, Aix-En-Provence, France*, 285–293.

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

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

Lalain, M., Ghio, A., Giusti, L., Robert, D., Fredouille, C., & Woisard, V. (n.d.). *Design and development of a speech intelligibility test based on pseudo-words in French: why and how?*

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.