



HAL
open science

Automatic modelling of perceptual judges in the context of head and neck cancer speech intelligibility

Sebastião Quintas, Mathieu Balaguer, Julie Mauclair, Virginie Woisard, Julien Pinquier

► **To cite this version:**

Sebastião Quintas, Mathieu Balaguer, Julie Mauclair, Virginie Woisard, Julien Pinquier. Automatic modelling of perceptual judges in the context of head and neck cancer speech intelligibility. International Journal of Language and Communication Disorders, 2024, pp.1–14. 10.1111/1460-6984.13004 . hal-04405083

HAL Id: hal-04405083

<https://hal.science/hal-04405083>

Submitted on 19 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

RESEARCH REPORT

Automatic modelling of perceptual judges in the context of head and neck cancer speech intelligibility

Sebastião Quintas¹  | Mathieu Balaguer^{1,2}  | Julie Mauclair¹ |
Virginie Woisard^{2,3}  | Julien Pinquier¹

¹IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France

²Laboratoire de NeuroPsychoLinguistique, UR 4156, Université de Toulouse, Toulouse, France

³IUC Toulouse, CHU Toulouse, Service ORL de l'Hôpital Larrey, Toulouse, France

Correspondance

Sebastião Quintas, IRIT Institut de Recherche en Informatique de Toulouse, SAMoVA Team, 118 Route de Narbonne, 31062 Toulouse Cedex 9, France.
Email: sebastiao.quintas@irit.fr

Funding information

Hospitals of Toulouse, and by the French National Research Agency, Grant/Award Number: ANR-18-CE45-0008; This project has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 766287

Abstract

Background: Perceptual measures such as speech intelligibility are known to be biased, variant and subjective, to which an automatic approach has been seen as a more reliable alternative. On the other hand, automatic approaches tend to lack explainability, an aspect that can prevent the widespread usage of these technologies clinically.

Aims: In the present work, we aim to study the relationship between four perceptual parameters and speech intelligibility by automatically modelling the behaviour of six perceptual judges, in the context of head and neck cancer. From this evaluation we want to assess the different levels of relevance of each parameter as well as the different judge profiles that arise, both perceptually and automatically.

Methods and Procedures: Based on a passage reading task from the Carcinologic Speech Severity Index (C2SI) corpus, six expert listeners assessed the voice quality, resonance, prosody and phonemic distortions, as well as the speech intelligibility of patients treated for oral or oropharyngeal cancer. A statistical analysis and an ensemble of automatic systems, one per judge, were devised, where speech intelligibility is predicted as a function of the four aforementioned perceptual parameters of voice quality, resonance, prosody and phonemic distortions.

Outcomes and Results: The results suggest that we can automatically predict speech intelligibility as a function of the four aforementioned perceptual parameters, achieving a high correlation of 0.775 (Spearman's ρ). Furthermore, different judge profiles were found perceptually that were successfully modelled automatically.

Conclusions and Implications: The four investigated perceptual parameters influence the global rating of speech intelligibility, showing that different judge profiles emerge. The proposed automatic approach displayed a more uniform

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *International Journal of Language & Communication Disorders* published by John Wiley & Sons Ltd on behalf of Royal College of Speech and Language Therapists.

profile across all judges, displaying a more reliable, unbiased and objective prediction. The system also adds an extra layer of interpretability, since speech intelligibility is regressed as a direct function of the individual prediction of the four perceptual parameters, an improvement over more black box approaches.

KEYWORDS

speech intelligibility, pathological speech, automatic speech processing, speaker embeddings, head and neck cancer

WHAT THIS PAPER ADDS

What is already known on this subject

- Speech intelligibility is a clinical measure typically used in the post-treatment assessment of speech affecting disorders, such as head and neck cancer. Their perceptual assessment is currently the main method of evaluation; however, it is known to be quite subjective since intelligibility can be seen as a combination of other perceptual parameters (voice quality, resonance, etc.). Given this, automatic approaches have been seen as a more viable alternative to the traditionally used perceptual assessments.

What this study adds to existing knowledge

- The present work introduces a study based on the relationship between four perceptual parameters (voice quality, resonance, prosody and phonemic distortions) and speech intelligibility, by automatically modelling the behaviour of six perceptual judges. The results suggest that different judge profiles arise, both in the perceptual case as well as in the automatic models. These different profiles found showcase the different schools of thought that perceptual judges have, in comparison to the automatic judges, that display more uniform levels of relevance across all the four perceptual parameters. This aspect shows that an automatic approach promotes unbiased, reliable and more objective predictions.

What are the clinical implications of this work?

- The automatic prediction of speech intelligibility, using a combination of four perceptual parameters, show that these approaches can achieve high correlations with the reference scores while maintaining a certain degree of explainability. The more uniform judge profiles found on the automatic case also display less biased results towards the four perceptual parameters. This aspect facilitates the clinical implementation of this class of systems, as opposed to the more subjective and harder to reproduce perceptual assessments.

INTRODUCTION

Loss of speech intelligibility is commonly found in conditions that affect the voice and vocal tract, such as head and neck cancer (HNC) and neurodegenerative diseases with

dysarthria or dysphonia symptoms. In the case of HNC, major functional repercussions on the upper aerodigestive tract (breathing, swallowing and phonation/speech) are likely to appear, hence a functional impairment at communication level is expected, impacting the patient's quality

of life (de Graeff et al., 2000). In order to quantify this loss of communication ability experienced, the perceptual evaluation of speech intelligibility is one of the most common methods of assessment used clinically (Balaguer et al., 2019; Dwivedi et al., 2009; Pommée et al., 2021a, 2021b).

Intelligibility is defined as the proportion of understood speech (Keintz et al., 2007) or the correctly transcribed word rate (Hustad, 2008; Pommée et al., 2021a, 2021b). At a higher level (i.e., words and sentences), syntactic and semantic contextual cues will enable the listener to reconstruct the information conveyed by the speech. These compensation phenomena, which require a correct intelligibility that is, a correct acoustic-phonetic decoding, is one of the elements responsible for comprehensibility in the context of degraded or pathological speech (Ghio et al., 2012; Hustad, 2008). Given this, despite the different ways of measuring speech intelligibility, the human component of this assessment can be seen as an inherently subjective process (Fex, 1992; Kent & Kim, 2003; Klopfenstein, 2009).

In clinical and research-related contexts, speech intelligibility can be evaluated in many different ways. Transcription has been characterized as a less subjective intelligibility measure (Miller, 2013) and involves the listener's direct transcription of the speaker's message word for word. Furthermore, the transcription is then compared to the target, and the resulting percentage of correctly transcribed words can serve as an intelligibility measure. Perceptual parameters such as prosody, phonation ability or voice quality are known to be intertwined with speech intelligibility (Balaguer et al., 2021; Bodt et al., 2002), and can be commonly found in some protocols as key parameters to assess in order to issue the intelligibility estimation (Sussman & Tjaden, 2012). The subjectivity comes into play in the different levels of relevance that each one of these parameters has (Miller, 2013). For example, some health practitioners can perceive prosody as a fundamental parameter for speech intelligibility while other practitioners can almost neglect it. Despite the heavy contribution to the overall subjectivity of intelligibility that these parameters have, they can also be seen as a roadmap to individual judge profiles. By understanding which parameters are more relevant to each judge, we can provide not only an explainable measure, but also a thorough understanding of the process behind the ratings proposed by different judges. Despite its efficiency, this process comes across as time-consuming. More subjective assessments, such as scaling tasks where judges are asked to rate intelligibility on a given scale (e.g., 0–10 rating, visual analog scales, etc.) are typically considered faster and comparable to transcription methods (Stipancic et al., 2016; Tjaden et al., 2014), however, at the expense of objectivity. On the other hand, perceptual evaluations are known to be very time-consuming, biased and variant, since the evaluation can be conditioned on, for example, patients previously assessed by the

same therapist (Fex, 1992). Different schools of thought can also affect interrater reliability, showcasing how hard to reproduce and subjective these measures can be, especially on mid-to-low intelligibility cases (Bunton et al., 2007). Due to the biased nature and low reproducibility of these scores, the development of an automatic assessment that is able to output unbiased and reproducible intelligibility measures becomes a plausible and interesting alternative to perceptual assessments (Balaguer et al., 2019). This aspect becomes relevant in the advent of creating automatic approaches that predict speech intelligibility, since a better understanding of data that are subjective by nature can provide added context towards high performance interpretable systems (Woisard, Astésano et al., 2021; Woisard Balaguer et al., 2021). Given this, the automatic modelling of speech intelligibility based on a combination of distinct perceptual parameters becomes of high interest.

From the literature, one can distinguish a variety of ways to automatically predict speech intelligibility. These methods can range from approaches such as regressing a score from the word error rate achieved by automatic speech recognition systems, known to underperform on severe patients (Christensen et al., 2012), to the extraction of relevant features from pathological speech, using automatic speech processing technologies (Fredouille et al., 2019; Middag et al., 2009; Quintas et al., 2023). Speaker embeddings, such as *i-vectors* or *x-vectors*, have also proved to be a viable alternative for intelligibility estimation. These fixed-dimension vectors aim to represent speaker-specific information in a vectorial sub-space, extracted from deep neural networks. These approaches have shown promising results when applied to the automatic assessment of pathological speech and automatic prediction of speech intelligibility (Laaridh et al., 2018; Quintas et al., 2020). In the context of the automatic prediction of subjective measures rated by a set of judges, systems that model the uncertainty associated with these measures become relevant. Given that high variances across judges can be found on some ratings, by devising a system that also takes variance into consideration instead of a general gold standard, a more robust system can normally be obtained (Chou & Lee, 2019; Fayek et al., 2016). Different ways to model uncertainty can be found in the literature, from modelling individual schools of thought (Rodrigues & Pereira, 2019) to using variance measures (e.g., SD) during training on some deep learning systems (Han et al., 2017; Rizos & Schuller, 2020). On the other hand, these systems that take in consideration the perception uncertainty normally do so in order to increase the performance metrics on the gold standard, instead of identifying and assessing individual judge profiles that can come across. Nevertheless, despite the recent progress in the automatic prediction of speech intelligibility, score interpretability is still a big issue (Quin-

tas et al., 2022) which typically impairs the widespread clinical usage of these technologies.

Since lack of agreement, variance and bias are common issues associated with perceptual measures, and since automatic systems typically lack explainability, the need for a better understanding of individual judge reasoning as well as interpretable automatic systems arises. Given that clinical measures such as speech intelligibility and speech disorder severity, evaluated with a 0–10 scale in the context of the present work, can be derived as a linear combination of other perceptual parameters, in the present work we intend to:

- illustrate the different perceptual judge profiles found when predicting speech intelligibility in HNC, through the means of perceptual parameters such as voice quality, resonance, prosody and phonemic distortions;
- propose an automatic approach that models each perceptual judge and predicts speech intelligibility as a combination of the same four perceptual parameters;
- show that our automatic approach not only achieves high base correlations, but also interpretable and more uniform intelligibility scores across all judges.

MATERIALS AND METHODS

C2SI corpus

The present work is based on the French head and neck cancer speech corpus Carcinologic Speech Severity Index (C2SI) (Woisard, Astésano et al., 2021). The corpus includes 87 patients that suffer oral cavity or oropharyngeal cancer and also 40 healthy speakers. Seven patients were recorded twice. Among the patients, 59% were men and the remaining 41% women. The mean age for the entire patient set was 65.8 years old (range 36–87). All cancer patients have undergone at least one cancer treatment, such as surgery, radiotherapy and/or chemotherapy. The applied cancer treatment lasted at least 6 months, after which the disorders are considered stable.

The patients, recorded in the context of the C2SI corpus, were recruited in the three main departments of Toulouse managing patients with HNC (ear, nose and throat department of the University Hospital, Cancerology department of the Institut Claudius Regaud (surgery and radiotherapy), maxillofacial surgery department of the Toulouse University Hospital). They were selected from the lists of carcinologic follow-up consultations of these three departments. These departments are part of the University Institute of Cancer in Toulouse (IUC-T) and associated with the unit of “Oncorehabilitation” which is located at the IUC-T Oncopole. These patients had to meet the following inclusion criteria:

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

Similarly, the criteria for non-inclusion were to present another source of speech disorders (e.g., stuttering) or to present cognitive or visual problems that are incompatible with the assessment protocol design. Table 1 presents a description of the study population.

All speakers were asked to record a different set of spoken tasks such as sustained vowels, picture description, spontaneous speech, passage reading and isolated pseudo-words. The speakers were comfortably seated in an anechoic room in front of a computer, which was used to visually display instructions and record the corpus. All speakers were recorded individually. For some tasks, the instructions were also produced with an auditory modality (e.g., pseudo-words).

The recordings were made with a Neumann TLM 102 Cardioid Condenser Microphone connected to a FOSTEX digital recorder. The sampling rate was 48 kHz, which facilitates the downsampling to 16 kHz, usually used in automatic speech processing. In this study, the main focus of attention was set towards the passage reading task, abbreviated to ‘LEC’, for ‘lecture’. In the context of the C2SI corpus, all speakers were asked to read the first paragraph of *La chèvre de M. Seguin*, a tale by Alphonse Daudet well known and widespread in French clinical phonetics (Ghio et al., 2012).

The full LEC task is as follows:

“Monsieur Seguin n’avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon. Un beau matin, elles cassaient leur corde, s’en allaient dans la montagne, et là-haut le loup les mangeait. Ni les caresses de leur maître ni la peur du loup rien ne les retenait. C’était paraît-il des chèvres indépendantes voulant à tout prix le grand air et la liberté”.

For each speaker, a set of perceptual parameters were assessed before the evaluation of speech intelligibility, based on the independent perceptual evaluation of six different health professionals (five speech and language pathologists and one phoniatrician). The four contemplated parameters are:

- Voice quality analysis (VQ),
- Resonance (R),

TABLE 1 Description of the study population.

Individual sociodemographic data	Patients	Controls
Subjects, <i>n</i>	87	42
Age, years, missing data	1	5
Extreme values	36–87	30–79
Shapiro–Wilk test: <i>p</i> value	0.63	0.03
Mean (SD)	65.8 (9.6)	60.1(12.7)
Median (interquartile range)	66 (59–72)	62 (55–68)
Gender, <i>n</i> (%)		
Missing data	0	3
Male	51 (59)	18 (46)
Female	36 (41)	21 (54)
Clinical data		
Location, <i>n</i> (%)		
Missing data	0	
Oral cavity	35 (40)	
Oropharynx	52 (60)	
TNM classification “T” (tumour size), <i>n</i> (%)		
Missing data	0	
T1	11 (13)	
T2	33 (38)	
T3	12 (14)	
T4	31 (35)	
TNM classification “N” (nodes), <i>n</i> (%)		
Missing data	17	
N0	22 (32)	
N1	17 (24)	
N2	3 (4)	
N2a	5 (7)	
N2b	13 (19)	
N2c	5 (7)	
N3	5 (7)	
Treatment data		
Time since the end of treatment, months		
Missing data	0	
Extreme values	6–239	
Shapiro–Wilk test: <i>p</i> value	<0.001	
Median (interquartile range)	39(21–91)	
Tumour surgery, <i>n</i> (%)		
Missing data	0	
Yes	73(84)	
No	14(16)	
Node surgery, <i>n</i> (%)		
Missing data	0	
Yes	76(87)	
No	11(13)	

(Continues)

TABLE 1 (Continued)

Individual sociodemographic data	Patients	Controls
Radiotherapy, <i>n</i> (%)		
Missing data	0	
Yes	82(94)	
No	5(6)	
Chemotherapy, <i>n</i> (%)		
Missing data	0	
Yes	48(55)	
No	39(45)	

- Prosody (**P**),
- Phonemic distortions (**PD**).

Each parameter was rated between 0 and 3. No definition of these concepts was given to the experts. The higher the value, the higher the impairment of that specific parameter. Similarly, speech intelligibility was also rated. Each speaker was given a score between 0 and 10, the smaller the value, the less intelligible the speech.

Statistical analysis

In order to find the most relevant parameters for speech intelligibility, according to each perceptual and posterior respective automatic judge, two statistical methods were employed. The first method was a multivariate analysis (Good, 2005) that uses robust linear regression analyses as a top-down variable selection approach, where the explanatory perceptual parameters were chosen based on the significance of their coefficient (using the *p* value and the 95% confidence interval). The second method studied was a grid search system (Barbero Jiménez et al., 2007) that was employed to search for the optimal combination of weights given a set of constraints. In our particular context, each weight is associated with one of the four perceptual parameters aforementioned. Inspired by the works of (Bodt et al., 2002), that conducted a perceptual modelling of speech disorder severity through the means of four similar perceptual parameters, Equation (1) illustrates the way intelligibility can be regressed using these same parameters. The grid search is optimized according to the predicted speech intelligibility ($INT(J_n)$) and the perceptual speech intelligibility, rated by each perceptual judge. The different weights (σ_n) refer to the relevance of each perceptual parameter through the form of a percentage, therefore being bounded between 0.0 and 1.0, under the condition that their sum equals to 1.0.

$$INT(J_n) = 10 - (5/6)(\sigma_1 \mathbf{VQ}_{J_n} + \sigma_2 \mathbf{R}_{J_n} + \sigma_3 \mathbf{P}_{J_n} + \sigma_4 \mathbf{PD}_{J_n}) \quad (1)$$

Due to the nature of the four perceptual parameters, that have a different and inverted scale when compared to speech intelligibility ([0,3] opposed to [0,10], see Sub-section C2SI Corpus), the sum of the parameters and respective weights are scaled accordingly, so that the resulting intelligibility value is comprised between a [0,10] range.

Automatic prediction of speech intelligibility

In order to model the behaviour of each perceptual judge, an ensemble of automatic models was created. Figure 1 displays a global overview of the proposed system, where each model predicts speech intelligibility at judge level through the means of the individual prediction of the four perceptual parameters (VQ, R, P and PD). The mean of the individual predictions of each judge is used to obtain the final intelligibility score for each speaker. In our context, we define the gold standard as the final intelligibility score, provided by the mean of each judge's individual intelligibility prediction (see *INTG* label in Figure 1) The system was based on the work of (Quintas et al., 2020), which made use of the *x-vector* speaker embedding paradigm (Snyder, Garcia-Romero, McCree et al., 2018); for the automatic prediction of speech intelligibility.

X-vector speaker embeddings

X-vectors are deep neural network speaker embeddings that have seen a growing use in speaker recognition and paralinguistic tasks (Pappagari et al., 2020; Snyder, Garcia-Romero, McCree et al., 2018; Snyder, Garcia-Romero, Sell et al., 2018; Snyder et al., 2017). These embeddings aim to represent discriminative features between speakers. The extractor system is trained primarily for speaker

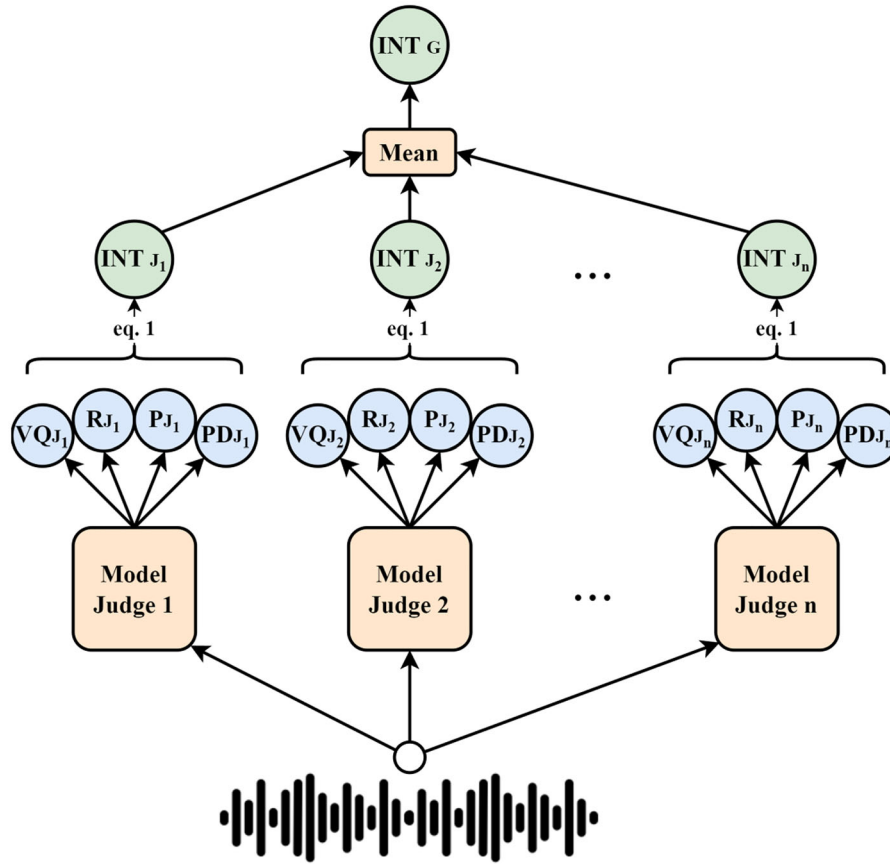


FIGURE 1 Global overview of the proposed system. The x -vectors are extracted from the segmented parts of a reading passage task (LEC), and then fed to an individual shallow neural networks that model each perceptual judge that regresses an intelligibility score. Abbreviations: INT, intelligibility; LEC, lecture; P, prosody; PD, phonemic distortions; R, resonance; VQ, voice quality. [Colour figure can be viewed at wileyonlinelibrary.com]

verification; therefore, utterances issued by the same speaker are closer in the vector space than utterances issued by different speakers. Besides their application on a variety of tasks, x -vectors have also outperformed other types of embeddings (e.g., i -vectors (Laaridh et al., 2018)) in the automatic prediction of speech intelligibility (Quintas et al., 2020).

In order to extract x -vectors, the open-source implementation present in the Kaldi toolkit¹ was used. The complete description of the x -vectors as well as the rationale behind the extraction network can be found in (Snyder et al., 2018). The embedding extractor works by first passing the speech signal through a block of time-delayed neural networks (TDNN) that operates on speech frames with a small temporal context centred at the current frame. Subsequent TDNN layers build on the temporal context of previous layers. A statistic pooling layer aggregates all frame-level outputs into a fixed-length dimension, which is then fed to a fully connected block. The x -vectors are extracted from

the affine component of the last fully connected layer and subsequently used as input features to a shallow neural network.

Shallow neural network

A shallow neural network was modelled to each judge, whose dimensions can be found on Table 2, as well as depicted on Figure 2. The loss function used takes in consideration all of the four perceptual parameters of each judge equally (VQ, R, P and PD), containing a total of 24 final parameters to optimize (6 judges \times 4 respective perceptual parameters), following a multi-task learning methodology (Zhang & Yang, 2018) where the proposed system learns different tasks simultaneously, hypothesizing that the synergy shared between similar tasks (i.e., same parameter prediction by each judge) conducts overall better automatic predictions.

TABLE 2 Shallow neural network outline used for each automatic judge.

Layer		Input × output
Shared layers	FC1	512 × 128
	FC2	128 × 64
Prediction layers	FC: Voice quality	
	FC: Resonance	
	FC: Prosody	64 × 1
	FC: Phonemic distortions	

Note: The first two layers correspond to the shared layers of each automatic judge while the remaining four correspond to the individual prediction layer of each one of the four perceptual parameters.

Abbreviation: FC, fully connected layers.

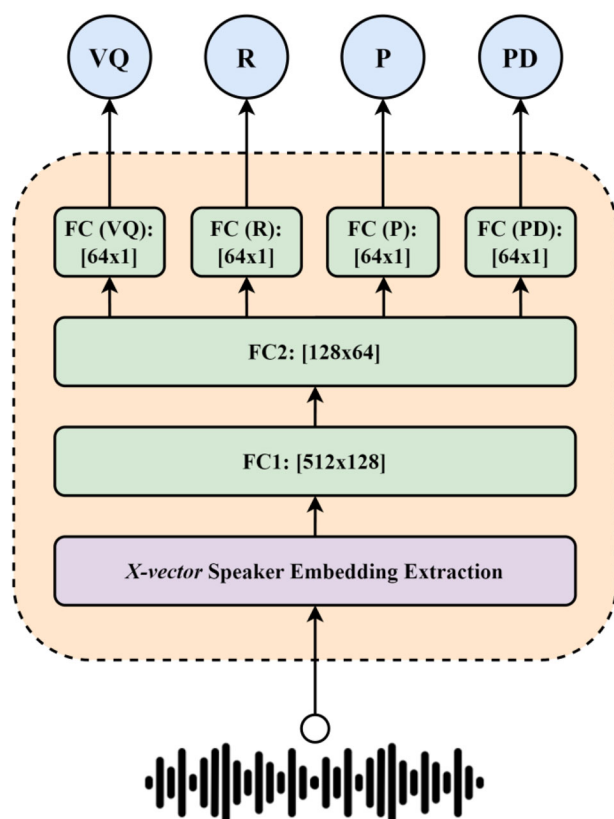


FIGURE 2 Diagram of the shallow neural network used to model each individual perceptual judge. Each network is trained to individually predict the four perceptual parameters. FC stands for fully connected layers, also known as layers where all the inputs from one layer are connected to every activation unit of the next layer.

Abbreviations: P, prosody; PD, phonemic distortions; R, resonance; VQ, voice quality.

[Colour figure can be viewed at wileyonlinelibrary.com]

EXPERIMENTS AND RESULTS

Judge-based automatic intelligibility modelling

In the present section, we will illustrate our experiments that correspond to the modelling of the judge set used to assess speech intelligibility. In this section we aim to show how we built our automatic model as well as to explore how the automatic modelling compares to the perceptual case.

Individual parameter modelling

As it was previously stated in Section Materials and Methods, an individual model was fit for each one of the six judges, that predicts the four perceptual parameters (VQ, R, P and PD). The set of models was optimized simultaneously. In order to train the proposed system, a 10-fold cross-validation scheme was implemented. Twenty-three controls of the C2SI corpus were added to the cross-validation folds as a data augmentation. These controls underwent the exact same procedure as the patients in terms of recorded tasks and perceptual evaluations (see Section C2SI Corpus). At each fold, 97 speakers were used for training, while the remaining 13 speakers were left for validation. The system was trained during 13 epochs, with a batch size of 32. A learning rate of 0.001 was used and the system was optimized with the Adam optimizer algorithm. A dropout of 0.1 was added between consecutive layers. No controls were used during the posterior evaluation of the system.

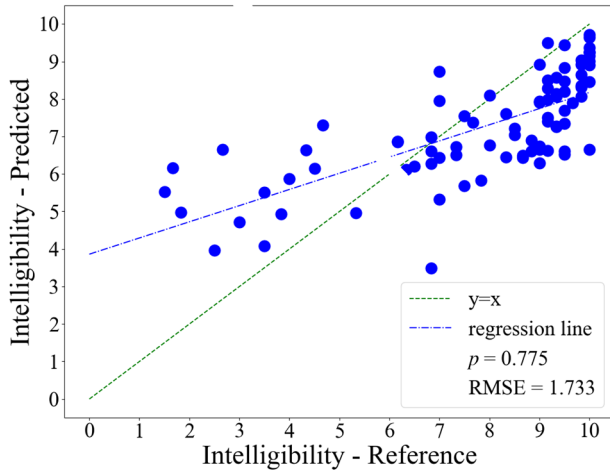


FIGURE 3 Results of the automatic prediction of speech intelligibility using the proposed methodology on the 85 patients. The intelligibility prediction in this plot corresponds to the global intelligibility given by the proposed set of automatic judges (see “INT Global” from Figure 1), also known as the gold standard. Abbreviations: INT, intelligibility; RMSE, root mean squared error. [Colour figure can be viewed at wileyonlinelibrary.com]

Intelligibility prediction while deducting individual judge profiles using relevant parameters

The Spearman’s correlation coefficient (ρ) and the root mean squared error (RMSE) were chosen to evaluate our system. The target scores used were the individual ratings of each judge on the four perceptual parameters of voice quality, resonance, prosody and phonemic distortions. The results, illustrated on Figure 2, present our gold standard prediction based on the automatic prediction of each one of the individual perceptual parameters of each judge, when all parameters of the Equation (1) share the same weight ($\sigma_{1.4} = 0.25$, uniform weights). The results suggest a high correlation ($\rho = 0.775$) and a small error ($RMSE = 1.733$), which can be found illustrated on Figure 3. Due to the low representation of patients with low intelligibility in the corpus, it was expected that the system would underperform in this particular case, specially given the nature of data-driven approaches that learn from previously seen (similar) samples. This aspect was evident on Figure 3, that presents less accurate predictions on these specific patients. A line obtained from the linear regression of our data points was added to the same figure to better illustrate this aspect, when compared to the line equivalent to a perfect prediction ($y = x$).

Since the system was trained using the four perceptual measures as the target of each judge (see Figure 1), it becomes relevant to assess how the corresponding automatic model for each human judge performs on these

TABLE 3 Correlation values of each judge’s perceptual parameters when compared to the reference values of those same individual parameters.

Judge	Correlation values (Spearman’s ρ)			
	VQ	R	P	PD
1	0.324	0.472	0.303	0.604
2	0.183	0.493	0.420	0.642
3	0.336	0.552	0.542	0.632
4	0.406	0.525	0.162	0.711
5	0.586	0.361	0.311	0.578
6	0.486	0.542	0.358	0.571
Mean	0.386	0.487	0.349	0.623

Abbreviations: P, prosody; PD, phonemic distortions; R, resonance; VQ, voice quality.

individual measures. Table 3 presents the results of this analysis, where each perceptual parameter is correlated to the predicted outcome at judge level. The results suggest a clear tendency towards the phonemic distortions being the perceptual parameter that achieves higher correlations for each speaker. From Table 3, we can also see that the voice quality and prosody parameters were the ones that achieved the lowest correlations. The correlation results between the individual parameters of voice, resonance, prosody and phonemic distortions, either perceptual or automatically predicted, and the reference speech intelligibility can be found on the Table A1 of the annex.

Four parameters versus speech intelligibility

An analysis to find the most relevant parameters for speech intelligibility was performed. Similarly to (Balaguer et al., 2021), a multivariate analysis was conducted. The relevant results of this multivariate analysis (see bold values of the Table 4) were used as a constraint for the posterior grid search analysis. Table 5 presents the correlation values obtained between the regressed intelligibility (see Equation 1) and the reference intelligibility score of each judge, on both the perceptual and automatic judges. Three different weight strategies were implemented. The uniform weights approach is based on an equal contribution between all four perceptual parameters ($\sigma_{1.4} = 0.25$). The constrained and general weights are both based on the grid search previously introduced, with the constrained weights being conditioned on the multivariate analysis (i.e., only the bold parameters of the Table 4 were taken in consideration per judge). The resulting weights from this grid search analysis can be found on the Table A1, present on the annex.

TABLE 4 Results of the robust linear regression analysis, used to find significant parameters that explain speech intelligibility, by listeners.

Judge	Multivariate analysis (Spearman's ρ)			
	VQ	R	P	PD
1	0.14	0.22	0.18	0.52
2	0.04	0.03	0.40	0.61
3	0.06	0.01	0.29	0.54
4	0.10	0.13	0.05	0.82
5	0.0	0.09	0.17	0.56
6	0.02	0.13	0.43	0.38

Note: Bold values represent significant correlations.

Abbreviations: P, prosody; PD, phonemic distortions; R, resonance; VQ, voice quality.

The correlations obtained behaved differently in the automatic approach than the perceptual case. While perceptually we can clearly see that the correlation increases on all judges when moving from the uniform weights to the constrained and later on to the general weights (see “Perceptual” columns of the Table 5), in the automatic approach the constrained approach promoted a drop in correlation (see “Automatic” columns of the Table 5). This aspect shows that the results from the multivariate analysis (illustrated on the Table 4) do not necessarily hold when shifting to automatic judges. Moreover, only small increases in correlation were observed on the automatic judges between the uniform and general weights. This aspect shows that the relevance levels of each of the four parameters are more uniform and less skewed when compared to the perceptual approach.

DISCUSSION

Main conclusions

In the present study, we proposed a way to automatically predict speech intelligibility as a combination of four perceptual parameters, based on their individual automatic prediction. The prediction obtained imply a reliable prediction of speech intelligibility, achieving a correlation of 0.775 (Spearman's) and an RMSE of 1.733. Furthermore, this aspect highlighted the individual judge profiles found, which made an analysis on individual schools of thought possible. We hypothesized that the four parameters studied were deeply connected to speech intelligibility. The results suggest that, despite speech intelligibility being related to those four perceptual parameters, the degrees of relevance of each parameter vary across different raters. Given this, two types of judge profiles were identified

and compared. The first is based on the perceptual measures obtained by the human judges, and the second on the respective automatic modellings. The results show that, despite both approaches following individual judge profiles, the relevance of each perceptual parameter is more evenly distributed in the automatic case when compared to the more skewed distribution of the perceptual case.

Concerning the reference perceptual measures which were assessed by the ensemble of six human judges, we found that both prosody and phonemic distortions played a more relevant level for speech intelligibility (see Table 4). When using Equation (1) to compute the intelligibility score based on the four perceptual parameters, we can see an increase in correlation, when adding more relevance to those parameters (see Table 5), Constrained Weights Perceptual column). Both grid search approaches promoted correlation gains on all judges when compared to a uniform approach where all parameters have the same relevance. This aspect shows that different profiles emerge, highlighting the subjectivity and bias associated with these clinical measures.

On the other hand, when modelling those same perceptual judges via our automatic approach, we can see that the profiles that emerge from each judge are different from the ones found previously. In this case, the weight distribution for those same four parameters appears to be more uniform for all judges, where all automatic judges take in consideration at least three parameters, contrary to what was witnessed on the perceptual case (see Table A2 on the annex). This aspect is validated by the correlation results obtained on the Table 5 (Automatic columns), that show only minor gains between the uniform and the general (optimal) weights. Contrastingly, by constraining the weights on the results of the multivariate analysis of the perceptual measures, the results tend to remain fairly similar, showing that the profiling made via the perceptual analysis does not necessarily hold in an automatic setting, at least compared to a uniform approach. We believe these differences are a product of the system's initial optimization, which aimed for equal prediction performance across all four parameters, as opposed to the perceptual judges that clearly favoured the phonemic distortions parameter. This aspect shows that despite the subjectivity witnessed among the different perceptual judge profiles, displayed by the different levels of relevance for each parameter, an automatic approach will promote a more uniform and objective way to predict speech intelligibility based on those four parameters. This objectivity is invaluable when applied in a clinical context, since it provides an explainable intelligibility score as a function of the four perceptual parameters, while negating individual schools of thought.

TABLE 5 Results of the robust linear regression analysis, used to find significant parameters that explain speech intelligibility, by listeners.

Judge	Correlation values (Spearman's ρ) INT (see Equation 1)					
	Uniform weights		Constrained weights		General weights	
	Perceptual	Automatic	Perceptual	Automatic	Perceptual	Automatic
1	0.828	0.613	0.839	0.616	0.851	0.630
2	0.774	0.699	0.849	0.700	0.852	0.728
3	0.836	0.613	0.879	0.623	0.879	0.633
4	0.767	0.726	0.813	0.679	0.863	0.742
5	0.648	0.490	0.784	0.452	0.784	0.539
6	0.750	0.637	0.792	0.627	0.810	0.649

Abbreviation: INT, intelligibility.

Perspectives

The different results obtained showed that out of all the perceptual measures studied, it was the PD parameter that had the biggest influence on intelligibility perceived by all the expert judges, being the only statistically significant parameter all the time. This aspect confirms the consensual intelligibility definition given by (Pommée et al., 2021a, 2021b) (Delphi study). Despite this, it may also be interesting to take an interest in the overall severity of the disorder (Balaguer et al., 2021) since it provides a broader notion than intelligibility, in which we could find different information and subsequent profiles at the level of the four parameters. An automatic modelling of speech disorder severity could then make it possible to highlight the role of the same four parameters in this concept, and thus adapt better the therapeutic strategies to the needs of the patients.

In the present work, we proposed the modelling of a set of measures that are subjective by nature, since no prior explanation of these measures was given to the set of judges before their assessments. The different profiles that emerge from this analysis display the subjectivity associated with the perceptual measures, where each judge takes in consideration different parameters for the evaluation of intelligibility. This subjectivity is considered one of the main impairments behind the clinical evaluation of speech intelligibility, among with judge bias, variance and lack of agreement. Moreover, we can see that both perceptual and automatic judges follow different approaches concerning the relevance of each perceptual parameter. Despite this, the results from the automatic modelling showed that we can predict speech intelligibility in a more objective way, where each automatic judge uses a more uniform approach concerning the relevance of each perceptual parameter.

The approach devised during the course of this work is generic, and therefore can be implemented in a variety of cases that involve the modelling of tasks that are subjective

by nature, and that require a set of judges. These tasks can range from, for example, sentiment analysis and opinion scoring to medical diagnosis from images or speech. Despite the approach being implemented in the context of professional judges, it can also be easily adapted to naive judges, which could provide relevant (automatic) clinical information concerning the perceived day-to-day communication of a given patient. The modelling of individual judges allows the attainment of not only the gold standard, but also the corresponding automatic score of each judge. This aspect provides extra clinical information and flexibility, as it can serve not only as a set of external opinions to a given practitioner (Balaguer et al., 2019), but also as a tool to label extensive quantities of subjective data (Haralabopoulos et al., 2020), which can be seen as highly relevant either in clinical contexts or research activities. The development of reliable and interpretable automatic systems allows the introduction of these systems in a clinical environment. A study on the direct impacts that a hybrid approach can have on the prediction of speech intelligibility, for example, mixing a set of automatic and perceptual judges, is an interesting lead for future work, as well as exploring the relevance of more objective and explainable automatic parameters on both perceptual and automatic scores.

CONCLUSIONS

In the present work, we studied the relationship between four perceptual parameters and speech intelligibility, through the means of an automatic modelling of a set of judges. Before predicting speech intelligibility, each human judge assessed the perceptual parameters of voice quality, resonance, prosody and phonemic distortions. A statistical analysis revealed that despite these four parameters being highly related to speech intelligibility, different judge profiles emerge, that attribute different levels of relevance to each parameter. An ensemble of automatic

models was then proposed to model each individual perceptual judge. The results suggest a high level of correlation ($\rho = 0.775$) when predicting the gold standard. Moreover, the same statistical analysis on the individual automatic judges also suggested individual judge profiles, however, this time with a more uniform relevance level found across the four parameters, instead of the skewed distribution previously found on the perceptual judges. The results suggest that the four perceptual parameters analysed are both highly relevant for speech intelligibility, both for the perceptual and automatic case. Furthermore, due to the more uniform relevance levels found on the automatic judges, we can consider the automatic approach more objective and interpretable.

ACKNOWLEDGEMENTS

This study was funded by the Hospitals of Toulouse, and by the French National Research Agency [RUGBI project, Grant ANR-18-CE45-0008]. Similarly, This project has also received funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 766287.

CONFLICT OF INTEREST STATEMENT

All authors declare that they have no conflict of interest.


DATA AVAILABILITY STATEMENT

Data and the database are available on request from the corresponding author.

PATIENT CONSENT STATEMENT

Each subject was told in advance of the purpose of this study and was given an information sheet. Subjects gave their written informed consent (non-opposition).

ORCID

Sebastião Quintas  <https://orcid.org/0000-0002-8693-9638>

Mathieu Balaguer  <https://orcid.org/0000-0003-1311-4501>

Virginie Woisard  <https://orcid.org/0000-0003-3895-2827>

Note

¹<https://github.com/kaldi-asr/kaldi>

REFERENCES

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 Oto-rhino-laryngology, Head and Neck Diseases*, 136(5), 355–359.

Balaguer, M., Pommée, T., Farinas, J., Pinquier, J. & Woisard, V. (2021) Paramètres perceptifs expliquant la sévérité du trouble de

parole mesurée automatiquement en cancérologie ORL. *Rééducation orthophonique, Ortho édition, Chapitre : "De l'exploration à la prise en soins de la voix chez l'adulte : données actuelles.. sur la voie des voix"*, 286, 1–13.

- Balaguer, M., Pommée, T., Farinas, J., Pinquier, J., Woisard, V. & Speyer, R. (2019) Effects of oral and oropharyngeal cancer on speech intelligibility using acoustic analysis: systematic review. *Head & Neck*, 41(1), 111–130.
- Barbero Jiménez, Á., Lázaro, J.L. & Dorronsoro, J.R. (2007) Finding optimal model parameters by discrete grid search. *AINSC Advances in Soft Computing*, 77(13), 2824–2832.
- Bodt, M., Huici, M. & Heyning, P. (2002) Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of Communication Disorders*, 35(3), 283–292.
- Bunton, K., Kent, R.D., Duffy, J.R., Rosenbek, J.C. & Kent, J.F. (2007) Listener agreement for auditory-perceptual ratings of dysarthria. *Journal of Speech, Language and Hearing Research*, 50(6), 1481–1495.
- Chou, H.C. & Lee, C.C. (2019) Every rating matters: joint learning of subjective labels and individual annotators for speech emotion classification. *Proceedings of ICASSP*, pp. 5886–5890.
- Christensen, H., Cunningham, S., Fox, C., Green, P. & Hain, T. (2012) A comparative study of adaptive, automatic recognition of disordered speech. *Proceedings of Interspeech*, pp. 1776–1779.
- de Graeff, A., de Leeuw, R.J., Ros, W.J., Hordijk, G.-J., Blijham, G.H. & Winnubst, J.A. (2000) Long-term quality of life of patients with head and neck cancer. *The Laryngoscope*, 110(1), 98–106.
- Dwivedi, R.C., Kazi, R.A., Agrawal, N., Nutting, C., Clarke, P.M., Kerawala, C.J., Rhys-Evans, P. & Harrington, K.J. (2009) Evaluation of speech outcomes following treatment of oral and oropharyngeal cancers. *Cancer Treatment Reviews*, 35(5), 417–424.
- Fayek, H.M., Lech, M. & Cavedon, L. (2016) Modeling subjectiveness in emotion recognition with deep neural networks: ensembles vs soft labels. *Proceedings of the International Joint Conference on Neural Networks*, pp. 556–570.
- Fex, S. (1992) Perceptual evaluation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 6(2), 155–158.
- Fredouille, C., Ghio, A., Laaridh, I., Lalain, M. & Woisard, V. (2019) Acoustic-phonetic decoding for speech intelligibility evaluation in the context of head and neck cancers. *International Congress of Phonetic Sciences (ICPhS)*, pp. 3051–3055.
- Ghio, A., Pouchoulin, G., Teston, B., Pinto, S., Fredouille, C., De Looze, C., Robert, D., Viallet, F. & Giovanni, A. (2012) How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers? *Speech Communication*, 54(5), 664–679.
- Good, P. (2005) Multivariate analysis. *Permutation, parametric and bootstrap tests of hypotheses*. 169–188.
- Han, J., Zhang, Z., Schmitt, M., Pantic, M. & Schuller, B. (2017) From hard to soft: towards more human-like emotion recognition by modelling the perception uncertainty. *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 890–897, <https://doi.org/10.1145/3123266.3123383>
- Haralabopoulos, G., M Tsikandilakis, M.T. & McAuley, D. (2020) Objective assessment of subjective tasks in crowdsourcing applications. *Proceedings of the LREC 2020 Workshop on Citizen Linguistics in Language Resource Development*, pp. 15–25.
- Hustad, K.C. (2008) The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of Speech, Language and Hearing Research*, 51(3), 562–573.

- Keintz, C.K., Bunton, K. & Hoit, J.D. (2007) Influence of visual information on the intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology*, 16(3), 222–234.
- Kent, R.D. & Kim, Y. (2003) Toward an acoustic typology of motor speech disorders. *Clinical Linguistics & Phonetics*, 17(6), 427–445.
- Klopfenstein, M. (2009) Interaction between prosody and intelligibility. *International Journal of Speech-Language Pathology*, 20(4), 326–331.
- 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 Interspeech*, pp. 2943–2947.
- Middag, C., Martens, J.P., Nuffelen, G.V. & Bodt, M.D. (2009) Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Advances in Signal Processing*, ArticleID 629030. <https://doi.org/10.1155/2009/629030>
- Miller, N. (2013) Measuring up to speech intelligibility. *Language & Communication Disorders*, 48(6), 601–612.
- Pappagari, R., Wang, T., Villalba, J., Chen, N. & Dehak, N. (2020) X-vectors meet emotions: a study on dependencies between emotion and speaker recognition. *Proceedings of ICASSP*, pp. 7169–7173.
- Pommée, T., Balaguer, M., Mauclair, J., Pinquier, J. & Woisard, V. (2021a) Intelligibility and comprehensibility: a Delphi consensus study. *International Journal of Language & Communication Disorders*, 57(1), 21–41.
- Pommée, T., Balaguer, M., Mauclair, J., Pinquier, J. & Woisard, V. (2021b) Assessment of adult speech disorders: current situation and needs in French-speaking clinical practice. *Logopedics Phoniatrics Vocology*, 47(2), 92–108.
- Quintas, S., Abad, A., Mauclair, J., Woisard, V. & Pinquier, J. (2023) Towards reducing patient effort for the automatic prediction of speech intelligibility in head and neck cancers. *Proceedings of ICASSP*, pp. 1–5.
- Quintas, S., Mauclair, J., Woisard, V. & Pinquier, J. (2020) Automatic prediction of speech intelligibility based on x-vectors in the context of head and neck cancer. *Proceedings of Interspeech*, pp. 4076–4980.
- Quintas, S., Mauclair, J., Woisard, V. & Pinquier, J. (2022) Automatic assessment of speech intelligibility using consonant similarity for head and neck cancer. *Proceedings of Interspeech*, pp. 3608–3612.
- Rizos, G. & Schuller, B.W. (2020) Average jane, where art thou?—recent avenues in efficient machine learning under subjectivity uncertainty. *IPMU International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 42–55.
- Rodrigues, F. & Pereira, F. (2019) Deep learning from crowds. *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1611–1618.
- Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D. & Khudanpur, S. (2018) Spoken language recognition using x-vectors. *Proceedings of Interspeech*, pp. 105–111.
- Snyder, D., Garcia-Romero, D., Povey, D. & Khudanpur, S. (2017) Deep neural network embeddings for text-independent speaker verification. *Proceedings of Interspeech*, pp. 999–1003.
- Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. & Khudanpur, S. (2018) X-vectors: robust DNN embeddings for speaker recognition. *Proceedings of ICASSP*, pp. 5329–5333.
- Stipancic, K.L., Tjaden, K. & Wilding, G. (2016) Comparison of intelligibility measures for adults with Parkinson's disease, adults with multiple sclerosis, and healthy controls. *Journal of Speech, Language and Hearing Research*, 59(2), 230–238.
- Sussman, J.E. & Tjaden, K. (2012) Perceptual measures of speech from individuals with Parkinson's disease and multiple sclerosis: intelligibility and beyond. *Journal of Speech Language and Hearing Research*, 55(4), 1208–1219.
- Tjaden, K., Kain, A. & Lam, J. (2014) Hybridizing conversational and clear speech to investigate the source of increased intelligibility in speakers with parkinson's disease. *Journal of Speech, Language and Hearing Research*, 57(4), 1191–1205.
- Woisard, V., Astésano, C., Balaguer, M., Farinas, J., Fredouille, C., Gaillard, P., Ghio, A., Giusti, L., Laaridh, I., Lalain, M., Lepage, B., Mauclair, J., Nocaudie, O., Pinquier, J., Pouchoulin, G., Puech, M., Robert, D. & Roger, V. (2021) C2SI corpus: a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers. *Language Resources and Evaluation*, 55, 173–190, <https://doi.org/10.1007/s10579-020-09496-3>
- Woisard, V., Balaguer, M., Fredouille, C., Farinas, J., Ghio, A., Lalain, M., Puech, M., Astesano, C., Pinquier, J. & Lepage, B. (2021) Construction of an automatic score for the evaluation of speech disorders among patients treated for a cancer of the oral cavity or the oropharynx: the carcinologic speech severity index. *Head & Neck*, 44(1), 71–88.
- Zhang, Y. & Yang, Q. (2018) An overview of multi-task learning. *National Science Review*, 5, 30–43. <https://doi.org/10.1093/nsr/nwx105>

How to cite this article: Quintas, S., Balaguer, M., Mauclair, J., Woisard, V., Pinquier, J. (2023) Automatic modelling of perceptual judges in the context of head and neck cancer speech intelligibility. *International Journal of Language & Communication Disorders*, 1–14. <https://doi.org/10.1111/1460-6984.13004>

APPENDIX 1

ANNEX

TABLE A1 Results from the Grid Search analysis on the significant parameters that explain speech intelligibility, on both the perceptual evaluations (Per.) given by each judge, and also on his/hers corresponding automatic evaluation (Aut.).

Judge	Constrained approach								General approach							
	VQ		R		P		PD		VQ		R		P		PD	
	Per.Aut.	Per.	Aut.	Per.	Aut.	Per.	Aut.	Per.	Aut.	Per.	Aut.	Per.	Aut.	Per.	Aut.	
1	---	0.30	0.10	0.30	0.50	0.40	0.40	0.10	0.33	0.20	0.0	0.30	0.33	0.40	0.33	
2	---	---	---	0.33	0.40	0.66	0.60	0.10	0.20	0.0	0.40	0.30	0.0	0.60	0.40	
3	---	---	---	0.10	0.33	0.90	0.66	0.0	0.0	0.0	0.30	0.10	0.50	0.90	0.20	
4	---	---	---	---	---	1.0	1.0	0.20	0.20	0.20	0.10	0.0	0.30	0.60	0.40	
5	---	---	---	---	---	1.0	1.0	0.0	0.0	0.0	0.30	0.0	0.60	1.0	1.0	
6	---	---	---	0.40	0.50	0.60	0.50	0.0	0.20	0.20	0.10	0.30	0.40	0.50	0.30	

Abbreviations: P, prosody; PD, phonemic distortions; R, resonance; VQ, voice quality.

TABLE A2 Correlations obtained between the perceptual intelligibility and the four perceptual parameters, obtained either perceptually or automatically.

Judge	Correlation values (Spearman's ρ) INT (see Equation 1)							
	Voice quality		Resonance		Prosody		Phonemic distortions	
	Perceptual	Automatic	Perceptual	Automatic	Perceptual	Automatic	Perceptual	Automatic
1	0.492	0.467	0.561	0.471	0.432	0.451	0.748	0.559
2	0.249	0.269	0.559	0.678	0.627	0.575	0.807	0.692
3	0.533	0.439	0.617	0.604	0.738	0.555	0.836	0.609
4	0.178	0.291	0.558	0.599	0.020	0.349	0.813	0.679
5	0.212	0.279	0.353	0.476	0.323	0.387	0.784	0.452
6	0.222	0.269	0.575	0.574	0.620	0.513	0.732	0.585

Abbreviation: INT, intelligibility.