Towards Reducing Patient Effort for the Automatic Prediction of Speech Intelligibility in Head and Neck Cancers

Sebastião Quintas, Alberto Abad, Julie Mauclair, Virginie Woisard, Julien Pinquier

ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094921

TOWARDS REDUCING PATIENT EFFORT FOR THE AUTOMATIC PREDICTION OF SPEECH INTELLIGIBILITY IN HEAD AND NECK CANCERS

Sebastião Quintas[1], *Alberto Abad*[2], *Julie Mauclair*[1], *Virginie Woisard*[3,4], *Julien Pinquier*[1]

[1]IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
[2]INESC-ID, Instituto Superior Técnico, Lisbon, Portugal
[3]IUC Toulouse, CHU Toulouse, Service ORL de l'Hôpital Larrey, Toulouse, France
[4]Laboratoire de NeuroPsychoLinguistique, UR 4156, Université de Toulouse, Toulouse, France

ABSTRACT

The automatic prediction of speech intelligibility can be seen as a growing and relevant alternative to the perceptual evaluations used clinically, which are known to be biased, variant and subjective. We propose an automatic way to regress an intelligibility score based on a recurrent model with a self-attention mechanism. This approach not only presented a high correlation of 0.87 when applied to a pseudo-word task designed for head and neck cancers, but also a significant decrease in error of more than 50%, when compared to previous approaches. Moreover, we have also studied the reliability of the same system when operating with smaller amounts of data at inference time. The results suggest that we can reduce the linguistic sample size to only 30% of the full sample, without losing performance. This aspect validates the reliability of using a smaller subset of data when predicting intelligibility, which can be extremely useful to prevent patient's fatigue by creating smaller batteries of clinical exams.

***Index Terms*—** Pathological Speech, Speech Intelligibility, Deep Learning, Head and Neck Cancer

1. INTRODUCTION

Loss of speech intelligibility is commonly found in the post-treatment of conditions that affect the vocal tract, such as head and neck cancer (HNC) and neurodegenerative diseases with dysarthria symptoms. 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 [1]. Since an early tracking and diagnosis are usually correlated to a better prognosis, due to the progressive and timed implementation of post-treatment measures, the perceptual evaluation of speech intelligibility has long been the most common method of disordered speech assessment. The perceptual assessments performed clinically normally include extensive recording tasks, that can become highly wearing on patients, namely the ones with severe speech impairments. Patient's fatigue is a common reason to leave some of these speech tasks incomplete. Due to this, it becomes relevant to devise more targeted and less extensive batteries of exams. Furthermore, these measures are known to be highly variable, biased and subjective, since the evaluation can be conditioned on several aspects such as prior knowledge of the task being issued (e.g. passage reading tasks), earlier assessments or also being conditioned on *a priori* knowledge of the patients. This aspect greatly affects reproducibility [2]. The usage of tasks based on pseudo-words and pseudo-sentences can mitigate the bias associated to the prior knowledge of the phonetic content being issued [3], however the majority of the issues that arise from a perceptual assessment still remain similar.

An automatic approach can be seen as an interesting alternative to perceptual measures, that could tackle the aforementioned issues. These approaches can promote more reliable, objective and reproducible scores [4, 5]. In order to automatically predict speech intelligibility, one can use approaches that regress a score from the word error rate obtained via an ASR (automatic speech recognizer), however this type of approach tends to underperform on speakers with severe speech impairments [6]. Other approaches make use of speech processing technologies based on the extraction of relevant features, such as MFCCs, filterbanks, speaker embeddings, etc. [7, 8, 9, 10]. Nevertheless, it is known that these type of data-driven systems tend to require significant amounts of data in order to operate efficiently [11]. The perceptual evaluation of speech intelligibility also differs from the one applied to the speech-in-noise paradigm [12]. While the definition of intelligibility may be similar to both, the perceptual decoding differs between the two. Traditional intelligibility predictors used in speech perception, such as STOI or E-STOI, require the usage of clean time-aligned signals, even in end-to-end approaches [13], which is unfeasible for pathological speech.

Given the fact that both patient's fatigue and the inherent subjectivity of the perceptual measures are recurrent clinical problems, and that data-driven automatic approaches tend to require larger amounts of data, in the present work we aim to:

- Present an automatic system that is capable of regressing objective and reproducible intelligibility measures (section 2), while achieving a high correlation with the perception-based measures. This is illustrated by our experiments in section 3, using a pseudo-word task from the French corpus of head and neck cancer.
- Prove that the same system can still achieve high correlations and accurate predictions when using significantly smaller amounts of data (section 4).

2. METHODOLOGY

The proposed methodology[1] is based on the usage of a recurrent model with self-attention, applied to a pseudo-word task (see section 3.1). We can divide the system into two distinct modules. In the first part, we make use of a recurrent model in order to obtain automatic scores for each pseudo-word. The second part corresponds

[1]https://github.com/Elquintas/Self-Attention-Speech-intelligibility

to the regression of an intelligibility score per speaker, based on the individual scores of each one's respective pseudo-words. Figure 1 illustrates our proposed method.

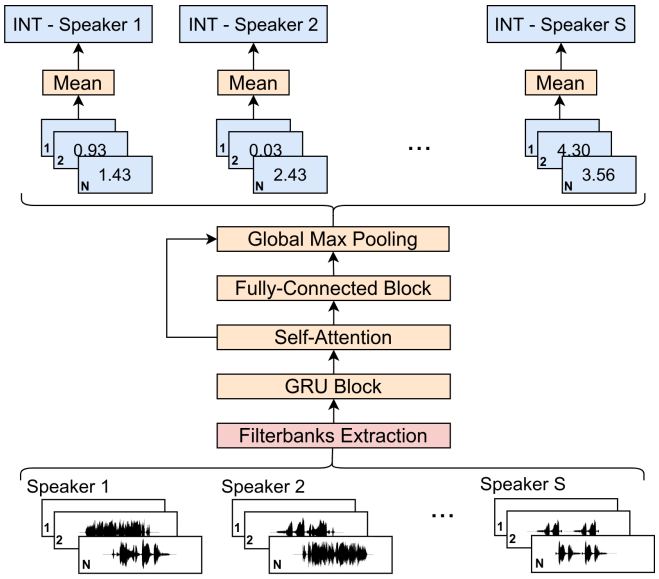

Fig. 1. Global overview of the proposed methodology. INT stands for the intelligibility score. N corresponds to the number of pseudo-words.

2.1. Modelization

The proposed system relies on a recurrent model with self-attention. Our system receives as input the individual audio files corresponding to each unique pseudo-word. From these audio files, we calculate on each window of 25ms with a 10ms stride, 40 filterbanks to be used as input features. For each pseudo-word, we use the intelligibility score obtained from its perceptual transcription as the target (see section 3.1). Our proposed model uses a bidirectional recurrent block, with three Gated Recurrent Unit (GRU) layers with an input dimension of 40 and a hidden dimension of 100. At the output of the block, a self-attention mechanism can be found, containing a single head. This mechanism allows the system to focus more on particular parts of the input file, while ignoring less relevant parts. After the attention mechanism, the utterance is passed to a set of three fully-connected layers, with a dimension of 100 units each and Rectified Linear Units (ReLUs) non-linearities. Finally, at the end of the system we have a Global Max Pooling layer used to obtain the individual score for each pseudo-word. Skip connections are added as shown in figure 1. By using the attention mechanism, we hypothesize that the system will automatically focus on and learn interesting interdependencies between the articulation of consecutive phonemes, that are known to be highly relevant for speech intelligibility [14]. The fully-connected block and max pooling aim to directly maximize the use of these learned interdependencies when predicting the final intelligibility score.

2.2. Automatic Regression of Speech Intelligibility

The second part of our proposed model corresponds to the regression of an intelligibility score for each speaker, as a function of the individual score of each pseudo-word. Instead of having the proposed model directly predicting the intelligibility score at speaker level, we opted to do the average of the set of each speaker's pseudo-words instead. This aspect promotes a more explainable measure, as a speaker's score can be traced down to the individual score of their pseudo-words. Interpretable systems are highly relevant clinically [15]. The usage of the average set of pseudo-words also gives us more flexibility to evaluate the importance of specific types of words in the automatic score.

3. EXPERIMENTS AND RESULTS

3.1. C2SI Corpus

In the present study, the focus of attention was set to the pseudo-word task recorded in the context of the French corpus of head and neck cancer (C2SI) [16]. Here, a set of speakers was invited to record each a set of 52 pseudo-words, that respect French phonotactic and orthographic rules. The pseudo-words are randomly drawn from a set of 89,346 possible forms, knowing that each list is, by construction, phonetically balanced. Due to the large amount of possible forms, the eventual repetition of a pseudo-word, despite being possible, is highly unlikely. The lists are balanced in the sense that they all contain the same number of each phoneme, but in different combinations (see [3] for further reference). Each pseudo-word follows the structure $C(C)_1V_1C(C)_2V_2$, where $C(C)_i$ is either a single consonant or a consonant group and V_i a single vowel. Each set of 52 words has a subset of 16 words with an occurrence of a double consonant (*d.c.*) at the beginning of the word, 16 words with *d.c.* in the middle and at least 26 words without *d.c.* The pseudo-words can have both occurrences of the *d.c.*, either at the beginning or at the middle, however in a smaller quantity (see table 1).

Table 1. Example of a set of 52 pseudo-words. Blue and violet represent the *d.c.* clusters.

banfou	**bl**eja	bou**ct**i	**br**im**pl**i	chessant	choniou	
clifant	cogu	**cr**impin	daillu	dinrant	**dr**edi	
fan**rs**i	**fl**in**rp**u	fouma	**fr**avi	gabi	**gl**unou	**gr**o**rv**o
guchin	joutu	juro	lanvin	le**rd**a	messo	mouco
nianlo	niejo	no**ks**a	nouillou	pa**st**u	pidant	
ploniou	**pr**ipin	**ps**ila	quiga	rinta	ru**rn**u	
san**vr**in	**sc**una	souquin	**sp**a**cl**ant	**st**icho	tan**gr**i	
tou**gz**u	**tr**a**dr**ou	vi**rj**ant	vumou	yainzi		
ya**lt**in	zebou	zouzant				

The perceptual intelligibility measure used in this study was obtained by averaging the individual transcription score of each pseudo-word by three distinct naive listeners, also known as perceived phonological deviation (PPD). This was used as reference throughout this work, and was calculated as a function of the distance between the transcribed word and the original one, taking into consideration a vowel and consonant matrix cost [17]. The distance between the transcribed pseudo-words and the canonical pronunciations was obtained using a Wagner-Fischer algorithm [18]. The perceptual scores range between 0, corresponding to the words perfectly pronounced, and 5, corresponding to unintelligible words. The final intelligibility score for each speaker was obtained by averaging the scores of their 52 pseudo-words. A set of 126 speakers was used, where 40 are controls and 86 patients.

3.2. System Training and Evaluation

In order to train the proposed system, a 10-fold cross validation scheme was employed. At each fold, a set of 113 speakers (patients and controls) are used for training, and the remaining 13 speakers for evaluation. For each fold, the system was trained during 200 epochs. A learning rate of 0.001 was used, with a polynomial decay until 0.0001 during the first 50 epochs. A batch size of 16 and the Adam optimizer were used during training as well. Spearman's Correlation (ρ) and the Root Mean Squared Error ($RMSE$) were chosen to evaluate our system. The target scores used were the perceptual intelligibility measures aforementioned in subsection 3.1, corresponding to the perceived phonological deviation.

The proposed approach is compared to two alternative systems. The first system is based on a Wagner-Fischer algorithm (previous baseline [19]) that aligns the canonical pronunciations with the results obtained from the automatic transcription of the pseudo-words. Due to the nature of the pseudo-words, the usage of a language model becomes invalidated. A text-constrained alignment is used instead, where each phoneme is represented by a three-state context-independent hidden markov model (HMM). Furthermore, after obtaining the phoneme boundaries, those are reconsidered by searching the most appropriate phoneme label among a set of 36 French phones while keeping the segment frontiers fixed (denoted as semi-constrained acoustic-phonetic decoding). In this manner, each speech segment available in the phoneme segmentation is confronted with the 36 three-state context-independent HMM obtained previously.

The second system compared to our approach is an intelligibility regression based on *x-vector* speaker embeddings [20]. Due to their nature, whose performance drops when using short file segments, the speaker embeddings were extracted from the file containing the entire set of pseudo-words, instead of extracting individual embeddings for each word. Similarly to [8], the *x-vectors* are used as features and fed to a shallow neural network that predicts speech intelligibility from the entire set of pseudo-words. A pre-trained model was used to extract the speaker embeddings. The shallow neural network was trained using a 10-fold cross-validation scheme.

The results, presented on table 2 and illustrated in figure 2, suggest a significant increase in correlation from 0.80 to 0.87, when compared to the latest previous system, and a drastic reduction on the RMSE values of more than 50 %, when compared to the Wagner-Fischer baseline.

Table 2. Comparison between two reference systems and our proposed approach.

	ρ	$RMSE$
Wagner-Fischer (baseline)	0.72	0.792
X-vector Speaker Embeddings	0.80	0.447
Recurrent Model with Self-Attention	**0.87**	**0.370**

3.3. Pseudo-Word Reduction

Clinically, the recording of 52 pseudo-words is not only time-consuming, but also highly wearing for the patients, namely the ones with more severe speech impairments. Given that patient's fatigue is not uncommon to happen in these recording sessions, which translates into leaving the recording session incomplete, it becomes thoroughly pertinent to evaluate how our proposed automatic approach behaves when predicting intelligibility with smaller quantities of data per speaker.

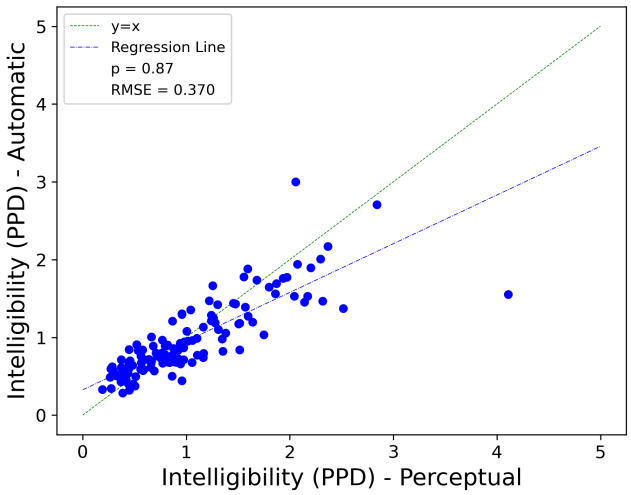

Fig. 2. Results of the automatic prediction of speech intelligibility using the Perceived Phonological Deviation (PPD).

Different sub-sets of pseudo-words have been investigated depending on the position and number of occurrences of *d.c.* [14] as shown in Table 3. For each reduced sub-set of pseudo-words, the final automatic score of every subject is computed as the mean of the scores of the sub-set, while the reference score still corresponds to the mean value of the 52 perceptual scores. The pseudo-word reduction only takes place during prediction and no new models were trained.

Given the results obtained with the two 16 words sets, with a *d.c.* at the beginning and middle respectively, we can only see minor changes both at correlation and at RMSE. These results corroborate the fact found in [14] for the perceptual measures, and therefore we can say that it is possible to use significantly smaller amounts of data whilst maintaining the reliability of the automatic measures promoted. For the remaining subsets of words, the results were similar, but they tend to accentuate the difference between pseudo-words with and without *d.c.* This aspect was particularly evident in the set without a *d.c.* (26 words), where besides having 10 more words than the two sets with *d.c.*, the correlation and error remain comparable. This analysis shows that pseudo-words with occurrences of *d.c.* are more relevant to obtain robust and reliable automatic intelligibility measures. This aspect was also illustrated by the results obtained in the last line of table 3, which displayed the results obtained with pseudo-words containing the two types of *d.c.*. Here a fairly decent correlation and RMSE values were obtained considering the small quantity of only five pseudo-words, that corresponds to less than 10% of the original quantity of pseudo-words used per speaker. We hypothesize that a larger subset of these words would outperform the other reduced lists, but we were limited by their small quantity in the corpus.

4. DISCUSSION

The results obtained suggest that it is possible to obtain high correlation values between the perceptual evaluations and automatic predictions. In addition, not only the correlation values increased greatly,

Table 3. Comparison between the scores obtained on the complete pseudo-word list and those of the reduced lists.

Model	Amount of pseudo-words used	ρ	$RMSE$	p-value
Wagner-Fischer Approach	52 (total)	0.72	0.792	< 0.05
X-vector Speaker Embeddings Approach		0.80	0.447	
Recurrent Model with Self-Attention	52 (total)	0.87	0.370	< 0.05
	16 with *d.c.* at the beginning	**0.85**	**0.370**	
	16 with *d.c.* at the middle	**0.85**	**0.375**	
	26 without *d.c.*	0.84	0.398	
	5 with *d.c.* at the beginning and middle	**0.79**	**0.413**	

but also the RMSE value drastically reduced to more than half compared to a prior approach [19].

Moreover, the analysis performed in subsection 3.3 showed that it is possible to remove words used in the global score and barely affect the reliability of the same automatic measures. This aspect becomes crucial clinically, where not only we can save precious time by recording fewer data, but also mitigate patient's fatigue. The usage of pseudo-words with *d.c.* was crucial to achieve these results, since it displayed a clear direction towards the best type of words to be used in these assessments. We believe that, these words tend to outperform their counterpart due to the larger phonetic content and key co-articulations between consonant and vowels and consecutive consonants. These co-articulations are an important marker for the post-operative assessment of head and neck cancers [21]. A larger presence of consonants is also hypothesized to be a good indicator of speech intelligibility [22, 23].

While keeping fewer pseudo-words, the results obtained with the *d.c.* subsets were encouraging, leaving the possibility for further reductions without largely affecting the scores. The creation of a reduced list, as well as working towards more interpretable scores through the means of the attention plots remain interesting leads for future work. The usage of a small number of words also becomes relevant clinically when compared to other automatic approaches that operate, for example, at sentence level [8], showing that high correlations can be achieved while using smaller amounts of data.

5. CONCLUSIONS

We proposed a new way to automatically predict speech intelligibility, based on a recurrent model with self-attention, for head and neck cancers. The proposed method achieved a high correlation value of 0.87 and a drastic reduction of more than 50% on the RMSE values, when compared to a previous approach based on the automatic transcription of the pseudo-word task. Moreover, the system was also studied on its reliability towards using smaller amounts of data. The results suggest that, we can significantly reduce the quantity of pseudo-words used given specific criteria whilst maintaining accurate predictions. The usage of pseudo-words with double-consonants was crucial for this reduction, showing that these words are indeed a viable indicator of speech intelligibility. This aspect becomes highly relevant in clinical practice, since it can not only help counter the variance and subjectivity associated with perceptual measures, but also alleviate patient's fatigue by recording smaller quantities of data.

6. ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 766287.

7. REFERENCES

[1] A. de Graeff, R. J. de Leeuw, W. J. Ros, G.-J. Hordijk, G. H. Blijham, and J. A. Winnubst, "Long-term quality of life of patients with head and neck cancer," *The Laryngoscope, Volume 110, Issue 1*, 2000.

[2] M. Balaguer, T. Pommée, J. Farinas, J. Pinquier, V. Woisard, and R. Speyer, "Effects of oral and oropharyngeal cancer on speech intelligibility using acoustic analysis: Systematic review," *Journal of the Sciences and Specialities of Head and Neck*, 2019.

[3] M. Lalain, A. Ghio, L. Giusti, D. Robert, C. Fredouille, and V. Woisard, "Design and development of a speech intelligibility test based on pseudowords in french: Why and how?" *Journal of Speech, Language and Hearing Research*, 2020.

[4] S. Fex, "Perceptual evaluation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 6, no. 2, pp. 155–158, 1992.

[5] C. Middag, *Automatic analysis of pathological speech*. Doctoral Dissertation: Ghent University, Department of Electronics and information systems, Ghent, Belgium, 2012.

[6] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," *Proceedings of Interspeech*, 2012.

[7] C. Middag, J.-P. Martens, G. V. Nuffelen, and M. D. Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, 2009.

[8] S. Quintas, J. Mauclair, V. Woisard, and J. Pinquier, "Automatic prediction of speech intelligibility based on x-vectors in the context of head and neck cancer," *Proceedings of Interspeech*, 2020.

[9] Y.-S. Lin and S.-C. Tseng, "Classifying speech intelligibility levels of children in two continuous speech styles," *Proceedings of ICASSP*, 2021.

[10] S. Quintas, J. Mauclair, V. Woisard, and J. Pinquier, "Automatic assessment of speech intelligibility using consonant similarity for head and neck cancer," *Proceedings of Interspeech*, 2022.

[11] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE Journal of Biomedical and Health Informatics*, 2017.

[12] A. H. Andersen, J. M. de Haan, Z.-H. Tan, and J. Jensen, "Non-intrusive speech intelligibility prediction using convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[13] M. B. Pedersen, M. Kolbæk1, A. H. Andersen, S. H. Jensen, and J. Jensen, "End-to-end speech intelligibility prediction using time-domain fully convolutional neural networks," *Proceedings of Interspeech*, 2020.

[14] A. Marczyk, A. Ghio, M. Lalain, M. Rebourg, C. Fredouille, and V. Woisard, "Have a cake and eat it too: Assessing discrimination performance of an intelligibility index obtained from a reduced sample size," *12th Conference on Language Resources and Evaluation*, 2020.

[15] W. K. Diprose, N. Buist, N. Hua, Q. Thurier, G. Shand, and R. Robinson, "Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator," *Journal of the American Medical Informatics Association, Volume 27, Issue 4*, 2020.

[16] C. Astésano, M. Balaguer, J. Farinas, C. Fredouillec, A. Ghio, P. Gaillard, L. G. I. Laaridh, M. Lalain, B. Lepage, and et al., "Carcinologic speech severity index project: A database of speech disorder productions to assess quality of life related to speech after cancer," *Language Resources and Evaluation Conference*, 2018.

[17] A. Ghio, M. Lalain, L. Giusti, C. Fredouille, and V. Woisard, "How to compare automatically two phonological strings: Application to intelligibility measurement in the case of atypical speech," *12th Conference on Language Resources and Evaluation (LREC)*, 2020.

[18] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *J. ACM*, vol. 21, pp. 168–173, 1974.

[19] C. Fredouille, A. Ghio, I. Laaridh, M. Lalain, and V. Woisard, "Acoustic-phonetic decoding for speech intelligibility evaluation in the context of head and neck cancers," *International Congress of Phonetic Sciences (ICPhS)*, 2019.

[20] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embedings for speaker recognition," *Proceedings of ICASSP*, 2018.

[21] G. Saravanan, V. Ranganathan, A. Gandhi, and V. Jaya, "Speech outcome in oral cancer patients - pre- and post-operative evaluation: A cross-sectional study," *Indian J Palliat Care*, 2016.

[22] Crevier-Buchman, V. J, M. S, and B. D, "Intelligibility of french consonants after partial supra-cricoid laryngectomy," *Revue de Laryngologie - Otologie - Rhinologie*, 2002.

[23] M. Fort, A. Martin, and S. Peperkamp, "Consonants are more important than vowels in the bouba-kiki effect," *Lang Speech*, 2015.