Measurement of speech intelligibility after oral or oropharyngeal cancer by an automatic speech recognition system

Mathieu Balaguer, Lucile Gelin, Virginie Woisard, Jérôme Farinas, Julien Pinquier

MEASUREMENT OF SPEECH INTELLIGIBILITY AFTER ORAL OR OROPHARYNGEAL CANCER BY AN AUTOMATIC SPEECH RECOGNITION SYSTEM

M. Balaguer[1,2], L. Gelin[1], V. Woisard[2,3], J. Farinas[1], J. Pinquier[1]

[1] IRIT, CNRS, Université Toulouse III, Toulouse, France
[2] Hôpital Larrey, CHU Toulouse, Toulouse, France
[3] Laboratoire Octogone-Lordat, Université Toulouse II, Toulouse, France

mathieu.balaguer@irit.fr, lucile.gelin@irit.fr, woisard.v@chu-toulouse.fr, jerome.farinas@irit.fr, julien.pinquier@irit.fr

Abstract:
Background: Speech intelligibility alteration is a frequent consequence of oral/oropharyngeal cancer. The development of automatic speech recognition (ASR) systems could overcome the limitations of perceptual speech assessment.
Objective: To predict speech intelligibility after treatment of oral or oropharyngeal cancer using scores from an ASR system.
Methods: Spontaneous speech of patients was recorded during a semi-structured interview. Six experts evaluated the subjects' intelligibility perceptually. An ASR system (TDNNf-HMM) trained on healthy adult speech and adapted to phoneme recognition was also used. Automatic scores were computed: phonemic scores, confidence scores. LASSO regression was used to select the parameters from the ASR system that best predicted intelligibility.
Results: Spontaneous speech of 25 patients was recorded. LASSO regression led to retain 3 parameters: number of sonants recognized per second, proportion of occlusives, and average confidence score of fricatives. These three parameters present a strong correlation (rs=0.91) with the perceptual score (expert panel). This automatically predicted score is stable and reliable (5-block cross-validation: rs= 0.90).
Conclusion: The use of ASR systems in the measurement of intelligibility in ENT oncology is promising. An optimization of these systems for pathological speech would open new perspectives for the determination of fine low-level speech deficits to adapt therapeutic objectives.
Keywords: **Speech, Automatic analysis, Oncology**

I. INTRODUCTION

Oral or oropharyngeal cancer alter speech abilities [1], in particular speech intelligibility. Intelligibility can be defined as the degree of accuracy with which the acoustic speech signal produced by a speaker is decoded by a listener in terms of "low-level" units (i.e., phonemes, phoneme groups, or syllables) [2].

Speech disorders are one indicator of intelligibility, and are mainly measured perceptually in clinical assessment [3]. Therapists quantify intelligibility using a variety of measurement tools, such as visual analog scales, Likert scale measures, or by measuring an error rate after transcription [2]. However, this standard perceptual evaluation has many limitations, particularly concerning its reliability. This measure is indeed judge-dependent, due to expertise effects or differences in internal referents [4]. Intra-individual variability effects are also involved: the same judge may assign different scores depending on the assessment context, the mental availability or habituation to pathological speech [5].

To overcome these biases, new tools for automatic instrumental speech assessment are being developed. They aim at extracting from the speech signal parameters for characterizing impairments [6]. These automatic and acoustic tools measure the quality of acoustic-phonetic decoding in a controlled speech context, such as text reading [7]. But few are applicable to spontaneous speech, due to a lack of a reference to which to compare the patient's speech – automatic alignment requiring prior manual transcription is too constraining to be applicable. Yet, this production context is the closest to the daily speech production [8] and needs to be investigated.

The objective is to predict speech intelligibility after treatment of oral or oropharyngeal cancer using scores from an automatic speech recognition system.

II. METHODS

This study is a cross-sectional observational study.

The study protocol was approved by the Committee for the Protection of Persons (CPP: Ouest IV, 19/02/2020, reference 11/20_3) within the framework of the ANR RUGBI project (https://www.irit.fr/rugbi, grant ANR-18-CE45-0008).

A. Participants

Patients coming for consultation or hospitalization in an ENT-oriented rehabilitation service or in an ENT consultation were included. Inclusion criteria were: being of legal age (at least 18 years old) and having been treated for oral or oropharyngeal cancer (surgical treatment and/or radiotherapy and/or chemotherapy, all tumors sizes) for at least six months (chronic and stable nature of the disorder). Exclusion criteria were: fatigable patients, associated pathology potentially responsible for speech or fluency disorders (e.g., stuttering, speech disorder from neurologic disease

B. Speech recordings

All subjects were recorded in a non-anechoic room, to be as close as possible to the usual clinical evaluations. No external or internal noise (such as air conditioning or ventilation. was to be perceptible in order not to disturb the quality of the recording. The speech samples were recorded on a ZOOM H4N Pro digital recorder (48 kHz sampling rate, 16-bit resolution, mono). The headset microphone (Thomann T.Bone HC 444 TWS) was placed 6 cm from the subject's mouth, positioned frontally below the level of the lower lip and at the level of the right labial commissure. For processing, the audio files were resampled to 16 kHz. The use of a Voice Activity Detector (WebRTC-VAD: https://github.com/wiseman/py-webrtcvad) was then used to isolate the subject's speech segments, excluding the examiner's speech segments.

To get a sample of spontaneous speech, the subjects were recorded during a semi-structured interview.

C. Speech analysis

A panel of expert listeners experienced in the evaluation of speech disorders was recruited to obtain a reference measure of intelligibility: one phoniatric physician and five speech therapists practicing in an ENT/oncology department.

The experts had to listen to the recording of the interview and to quantify the intelligibility on a scale from 0 (unintelligible) to 10 (totally intelligible). The baseline perceptual intelligibility score was the average of the scores given by the 6 judges.

The subjects' speech segments – determined by the Voice Activity Detector – were given as input to a TDNNf-HMM (factorized Time-Delay Neural Network - Hidden Markov Model [9]) ASR system. The model used in this study [10] was developed using the Kaldi toolkit [11] and adapted for phoneme recognition (Phone Error Rate=23.5% on a typical adult corpus [10]). The system was trained using the Common Voice online database: in French, the training corpus includes 148.9 hours of read text recordings, by 1,276 speakers. For decoding, in each 25 ms frame (with a 10 ms step), the phone closest to the acoustic features carried by the signal will be retained and associated with the corresponding phoneme (among 33 French phonemes). A confidence score is also associated to each recognized phoneme using a Minimum Bayes Risk method [12]. WIP (Word Insertion Penalty) and LMWT (Language Model Weights) have been set to their minimum value (WIP=0; LMWT=7) to obtain a raw output.

For each subject, 16 scores were calculated based on the system outputs (see details in Table 1).

D. Statistical analysis

The analyses were carried out using Stata 16.1 software (StataCorp. 2019. Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC.).

Due to the size of the study (n<30), the statistical tests used were nonparametric. In all analyses, a level of significance at 5% was chosen. For descriptive analysis, perceptual intelligibility and automatic scores were described by mean and median as indicators of central tendency, and by standard deviation, interquartile range, minimum and maximum values as indicators of dispersion. Correlations between intelligibility and automatic scores were analyzed using Spearman's correlation coefficients. Finally, the predictive process of automatic parameter selection was performed using LASSO regression (penalized regression).

III. RESULTS

A. Participants

Twenty-five patients were included (median age 67 years, IQR 12; 15 males, 10 females; oral cavity 14, oropharynx 10, both locations 1). 57.9% of patients were treated for a large tumor (T3 or T4). Surgical treatment was performed in 88% of cases (radiotherapy: 96%, chemotherapy: 60%, surgery and radiotherapy: 84%). The median time since the end of treatment was 40 months (range: 6-564 months).

B. Perceptual assessment of intelligibility (reference score)

Mean intelligibility was 6.87 (median: 7.17, range: 1.17-10). Inter-judge agreement was strong among the 6 expert judges: ICC=0.82 [0.72, 0.91].

C. Parameters from the ASR system: automatic scores

The 22 automatic scores were extracted for each subject (Table 1).

Table 1: Details of scores for the 22 automatic parameters from the ASR system

Parameter	Mean	SD	Median	IQR	Min. value	Max. value
Total of different phonemes recognized (*difphon*)	4.55	1.56	4.78	2.40	1.12	7.49
Number of phonemes recognized per second						
Total phonemes (*sumphons*)	29.20	5.57	32.00	3.00	5.00	32.00
Consonants (*csns*)	2.23	0.89	2.34	1.27	0.17	4.05
Occlusives (*occs*)	0.58	0.41	0.56	0.62	0.00	1.51
Fricatives (*fris*)	0.80	0.29	0.85	0.31	0.00	1.15
Sonants (*sonants*)	0.96	0.38	1.00	0.48	0.17	1.62
Nonsonants (*nonsonants*)	1.38	0.60	1.33	0.76	0.00	2.61
Vowels (*vows*)	2.22	0.65	2.33	1.06	0.96	3.32
Semi-consonants (*semicsns*)	0.11	0.08	0.11	0.11	0.00	0.36
Proportion of phonemes recognized among consonants						
Occlusives (*propocc*)	0.23	0.12	0.27	0.20	0.00	0.37
Fricatives (*propfri*)	0.36	0.14	0.34	0.16	0.00	0.78
Sonants (*propsonant*)	0.46	0.14	0.42	0.11	0.23	1.00
Nonsonants (*propnsonant*)	0.59	0.14	0.62	0.13	0.00	0.78
Proportion of phonemes recognized among vowels						
Nasal vowels (*propvnasal*)	0.18	0.10	0.17	0.09	0.05	0.44
Proportion of phonemes recognized among all phonemes						
Vowels (*propvow*)	0.51	0.09	0.49	0.04	0.43	0.85
Nasal phonemes (*propnasal*)	0.19	0.06	0.19	0.06	0.06	0.37
Confidence scores						
Overall (*conf*)	0.84	0.02	0.84	0.03	0.78	0.89
Consonants (*confc*)	0.87	0.04	0.88	0.03	0.76	0.93
Occlusives (*confo*)	0.87	0.07	0.90	0.09	0.72	0.95
Fricatives (*conff*)	0.88	0.04	0.88	0.05	0.79	0.93
Vowels (*confv*)	0.80	0.03	0.80	0.02	0.77	0.91
Semiconsonants (*confs*)	0.76	0.04	0.76	0.04	0.65	0.84

D. Parameters selection

Spearman's correlation coefficients are given as absolute values. Eight parameters (36%) show a high correlation with the baseline intelligibility score (rs≥0.70). Seven parameters (32%) showed moderate correlation (0.50≥rs>0.70). Details are shown in Fig. 1.

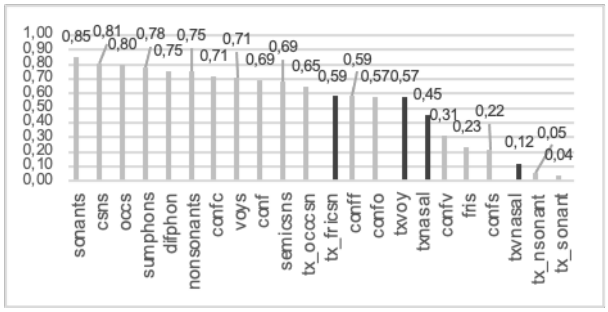

Fig. 1: Spearman's correlation coefficients between perceptual intelligibility and the 22 automatic scores (light grey: positive correlation coefficients, dark grey: negative ones).

Among the 22 automatic parameters, the LASSO regression allowed to select four parameters: the proportion of occlusives among consonants (*propocc*), the number of sonants per second (*sonants*), the average confidence score on fricatives (*conff*) and the number of occlusives per second (*occs*). An analysis of multicollinearity led to remove of the '*occs*' parameter (variance inflation factor = 7.17). The regression performed on the three remaining parameters explained 82.4% of the variance in intelligibility (R2), for a root mean squared error of 1.21. The predicted intelligibility is calculated as follows (1):

$$\text{intelligibility} = -0.073 + 4.982*\textit{sonants} + 6.188*\textit{propocc} + 0.851*\textit{conff} \quad (1)$$

The correlation between the perceptual intelligibility and the intelligibility predicted by the automatic parameters is rs=0.91 (p<0.001) (Fig. 2).

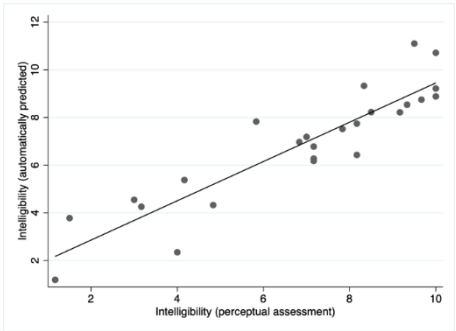

Fig. 2: Scatter plot between perceptual intelligibility and intelligibility by the three retained parameters

Cross-validation shows a strong correlation between the reference score and (i) intelligibility predicted by 5-block cross-validation (rs=0.90, p<0.001), (ii) intelligibility predicted by leave-one-out cross-validation (rs=0.90, p<0.001).

IV. Discussion

Intelligibility can be effectively predicted using three parameters from an ASR analysis of oral/oropharyngeal cancer speech. However, the results of this study can be considered preliminary due to the small sample size of subjects (n=25). The increase of the sample size would allow to conclude more strongly about the generalization and stability of these results.

The ASR system used is trained on typical (i.e., non-pathological) speech. Indeed, we wanted to measure a gap between healthy and pathological speech by targeting indicators of speech intelligibility. But one can wonder if training the system on pathological speech would allow to obtain more adapted acoustic models. In that case, if the acoustic models determined are more efficient (with a low Phone Error Rate in particular), the automatic scores calculated on the system outputs could perhaps allow to highlight finer deficits. Large corpora are necessary to train acoustic models that are relatively more stable given the pathological character of the speech. As no large French cancer speech corpus exists to date, transfer learning techniques can be used to adapt typical speech models to new corpora on relatively few data [13]. Specifically, it would be possible to adapt the current speech recognition system on other unused speech tasks in our corpus, such as sentences or text reading and pseudoword repetitions. Optimizing the quality of speech recognition could also involve the use of promising new ASR systems: the Listen, Attend and Spell (LAS) architectures [14], or Transformers [15]. These systems have been adapted to non-typical speech by Gelin [10], in this case children's speech. Their adaptation to oncologic speech would be relevant to study their performance.

ASR systems have multiple advantages in clinical evaluation: they are applicable to spontaneous speech, the scores are reliable, the required equipment is inexpensive, and the evaluation is fast. Thus, it remains relevant to explore the contributions of ASR for pathological speech analysis.

V. Conclusion

The use of ASR systems to assess intelligibility in ENT oncology is promising. An increase in sample size and analyses on optimization of these systems for pathological speech would open new perspectives for the determination of low-level speech deficits to adapt therapeutic objectives.

REFERENCES

[1] PA. Borggreven, IM. Verdonck-De Leeuw, MJ. Muller et al., "Quality of life and functional status in patients with cancer of the oral cavity and oropharynx: Pretreatment values of a prospective study", *European Archives of Oto-Rhino-Laryngology*, vol 264, pp. 651–657, 2007.

[2] KC. Hustad, "The Relationship Between Listener Comprehension and Intelligibility Scores for Speakers With Dysarthria", *Journal of Speech, Language, and Hearing Research*, vol. 51, pp. 562-573, 2008.

[3] T. Pommée, M. Balaguer, J. Mauclair, J. Pinquier, V. Woisard, "Assessment of adult speech disorders: current situation and needs in French-speaking clinical practice", *Logopedics Phoniatrics Vocology*, pp. 1–15, 2021.

[4] C. Kuo, K. Tjaden, "Acoustic variation during passage reading for speakers with dysarthria and healthy controls", *Journal of Communication Disorders*, vol 62, pp. 30–44, 2016.

[5] S. Fex, "Perceptual evaluation", *Journal of Voice*, vol 6, pp. 155-158, 1992.

[6] C. Middag, JP. Martens, G. Van Nuffelen, M. De Bodt, "Automated Intelligibility Assessment of Pathological Speech Using Phonological Features", *EURASIP Journal on Advances in Signal Processing*, 2009.

[7] M. Balaguer, T. Pommée, J. Farinas, J. Pinquier, V. Woisard, R. Speyer, "Effects of oral and oropharyngeal cancer on speech intelligibility using acoustic analysis: Systematic review", *Head and Neck*, vol 42, pp. 111-130, 2020.

[8] S. Knuijt, JG. Kalf, BGM. van Engelen, BJM. de Swart, ACH. Geurts, "The Radboud Dysarthria Assessment: Development and Clinimetric Evaluation", *Folia Phoniatrica et Logopaedica*, vol 69, pp. 143-153, 2017.

[9] D. Povey, G. Cheng, Y. Wang, et al., "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks", *Interspeech ISCA*, pp. 3743-3747, 2018.

[10] L. Gelin, M. Daniel, J. Pinquier, T. Pellegrini, "End-to-end acoustic modelling for phone recognition of young readers", *Available from: lalilo.com*, 2021.

[11] D. Povey, A. Ghoshal, G. Boulianne et al., "The Kaldi Speech Recognition Toolkit", *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[12] H. Xu, D. Povey, L. Mangu, J. Zhu, "Minimum Bayes Risk decoding and system combination based on a recursion for edit distance", *Computer Speech and Language*, vol 25, pp. 802-828, 2011.

[13] D. Wang, TF. Zheng, "Transfer learning for speech and language processing", *IEEE APSIPA*, pp. 1225-1237, 2015.

[14] W. Chan, N. Jaitly, Q. Le, O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition", *IEEE ICASSP*, pp. 4960-4964, 2016.

[15] L. Lu, C. Liu, J. Li, Y. Gong, "Exploring Transformers for Large-Scale Speech Recognition", *Interspeech ISCA*, pp. 5041-5045, 2020.