Automatic evaluation of spontaneous oral cancer speech using ratings from naive listeners

Halpern, Bence Mark; Feng, Siyuan; van Son, Rob; van den Brekel, Michiel; Scharenborg, Odette

DOI
10.1016/j.specom.2023.03.008

Publication date
2023

Document Version
Final published version

Published in
Speech Communication

Citation (APA)
Halpern, B. M., Feng, S., van Son, R., van den Brekel, M., & Scharenborg, O. (2023). Automatic evaluation of spontaneous oral cancer speech using ratings from naive listeners. *Speech Communication*, *149*, 84-97. https://doi.org/10.1016/j.specom.2023.03.008

Speech Communication 149 (2023) 84–97

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Automatic evaluation of spontaneous oral cancer speech using ratings from naive listeners

Bence Mark Halpern [a,b,c,*], Siyuan Feng [b], Rob van Son [a,c], Michiel van den Brekel [a,c], Odette Scharenborg [b]

[a] Netherlands Cancer Institute, Amsterdam, The Netherlands
[b] Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands
[c] University of Amsterdam, ACLC, Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Automatic speech evaluation
Pathological speech
Oral cancer

ABSTRACT

In this paper, we build and compare multiple speech systems for the automatic evaluation of the severity of a speech impairment due to oral cancer, based on spontaneous speech. To be able to build and evaluate such systems, we collected a new spontaneous oral cancer speech corpus from YouTube consisting of 124 utterances rated by 100 non-expert listeners and one trained speech-language pathologist, which we made publicly available. We evaluated the systems in two scenarios: a scenario where transcriptions were available (reference-based) and a scenario where transcriptions might not be available (reference-free).

The results of extensive experiments showed that (1) when transcriptions were available, the highest correlation with the human severity ratings was obtained using an automatic speech recognition (ASR) retrained with oral cancer speech. (2) When transcriptions were not available, the best results were achieved by a LASSO model using modulation spectrum features. (3) We found that naive listeners' ratings are highly similar to the speech pathologist's ratings for speech severity evaluation. (4) The use of binary labels led to lower correlations of the automatic methods with the human ratings than using severity scores.

1. Introduction

Oral cancer is a type of cancer where a tumour is located inside the oral cavity, most typically the tongue or floor of the mouth. Approximately 530,000 people get diagnosed with this condition every year worldwide (Shield et al., 2017), of which 53,000 in the USA alone (Foundation, 2019). To treat oral cancer, (part of) the tissues surrounding the tumour are removed during an operation, which subsequently affects the articulation abilities of the oral cancer patients. Moreover, problems with voice quality often occur due to the radiation treatment these patients may receive (Jacobi et al., 2010; Woisard et al., 2022; Balaguer et al., 2019). In certain cases, patients are able to learn articulatory compensation techniques to adjust for the lost tongue tissue (Ward and van As-Brooks, 2014). Learning these compensation techniques as part of speech therapy can alleviate speech problems in oral cancer speakers.

To evaluate the success of speech therapy for pathological speech, many types of perceptual speech evaluation measures can be used. Two of the most often used measures are intelligibility measures and voice quality measures. Intelligibility measures quantify the extent to which the speech could be transcribed by a naive or an expert listener (see for an application: Meyer et al. (2004)). Intelligibility measures are often preferred over other measures because the experimental setup is relatively easy, does not require expert listeners, and is often deemed sufficient for the evaluation of articulation disorders. However, intelligibility alone cannot measure all important aspects of pathological speech. For example, the speech could be completely intelligible while having a creaky voice quality. Evaluation of voice quality is usually done by speech-language pathologists (SLPs) through standardised questionnaires such as the Grade-Roughness-Breathiness-Asthenicity-Strain Scale (GRBAS) (Hirano, 1981) and the Consensus Auditory Perceptual Evaluation of Voice (CAPE-V) (Zraick et al., 2011). However, evaluations with SLPs are costly. The availability of a speech evaluation measure that gives a non-rigorous impression of the speech at both the level of intelligibility and voice quality provided by naive listeners (see Dagenais et al. (2006)) would therefore be quite useful.

Here we aim to develop a method to automatically evaluate the severity of the speech (in short: severity), which is defined as *the degree of the overall deterioration of the audible signal* (Balaguer et al., 2019), a

* Corresponding author at: Netherlands Cancer Institute, Amsterdam, The Netherlands.
E-mail addresses: b.halpern@nki.nl (B.M. Halpern), s.feng@tudelft.nl (S. Feng), r.v.son@nki.nl (R. van Son), m.vd.brekel@nki.nl (M. van den Brekel), o.e.scharenborg@tudelft.nl (O. Scharenborg).

https://doi.org/10.1016/j.specom.2023.03.008
Received 15 February 2022; Received in revised form 26 February 2023; Accepted 19 March 2023
Available online 25 March 2023

global measure that aims to quantify all speech severity properties - including intelligibility and voice quality.

There are a few existing methods that propose objective evaluation of oral cancer speech. Typically these methods focus on the intelligibility estimation of the speech (Windrich et al., 2008; Quintas et al., 2020; Bin et al., 2019) rather than voice quality estimation. Furthermore, these methods have only been validated with clean data and read speech. Evaluation on the basis of clean data and read speech is not necessarily ecologically valid, because, e.g., spontaneous speech is more indicative of the actual voice severity (Wolfe et al., 1995; Revis et al., 1999). Furthermore, an ideal objective evaluation method should be insensitive to channel noises and the type of recording devices.

Towards our ultimate aim to develop a more ecologically correct, robust, and objective automatic severity evaluation method for oral cancer speech, we collected an oral cancer speech dataset from YouTube with a wide variety of realistic speech conditions, which is more representative of oral cancer speakers' everyday speech than a read speech corpora. To the best of our knowledge, this corpus, which is an extension of our previous oral cancer dataset (Halpern et al., 2020), is the first publicly available oral cancer speech evaluation dataset. Two other datasets exist, a French and a German dataset; however, these are not publicly available (Windrich et al., 2008; Quintas et al., 2020).

The automatic speech severity evaluation task can be roughly described as a speech processing task where either one or multiple speech signals are fed into a processing function to obtain a single scalar number ($\hat{x} \in \mathbb{R}$) which is the estimate of the speech severity. This estimate can be compared against a ground truth severity score (x), which we obtained from both a speech-language pathologist and naive human listeners. The estimated speech severity is then correlated with the ground truth severity score, where a correlation of 1 indicates the perfect method for speech severity estimation.

The main aim of this work is to compare existing and new techniques for the automatic evaluation of the severity of the speech impairment due to oral cancer treatment (in short, oral cancer speech) to find the system that achieves the highest correlation with human ratings of the severity of the oral cancer speech. Therefore our main research question is the following: **RQ1: What automatic approach achieves the highest correlation with the ground truth severity scores for oral cancer speech severity evaluation?**

There are several paradigms for the objective evaluation of pathological speech. We divide these paradigms into two groups, i.e., reference-based and reference-free approaches. Reference-based methods use either a transcription of a speech signal (ASR-based methods) or a reference speech signal (comparison-based methods), while reference-free methods do not. In this work, we will compare both kinds of reference-based and several reference-free methods on the task of oral cancer speech severity evaluation.

ASR-based methods (Tripathi et al., 2020; Windrich et al., 2008; Maier et al., 2009) use the mistakes of speech recognisers to assess the speech intelligibility, which is often a good enough proxy for measuring severity of the speech. In other words, it is assumed that an ASR makes similar errors as an expert. Some transcription error measure (e.g., phoneme error rate, word error rate, Levenshtein distance) is used as an intelligibility estimate. ASR-based methods are often deemed the most useful methods because practitioners can directly inspect what words or phonemes the ASR system did not recognise. Their main disadvantage, though, is that a ground truth transcription of the pathological speech is required, which is often difficult to obtain, especially when the speech is unintelligible.

Comparison-based methods measure the distortion of a speech signal compared to a reference speech signal. These approaches originate from the speech enhancement (blind source separation) literature, where the distorted signal is a noised signal, which is compared to a clean signal (Vincent et al., 2006). Pathological speech, then, can be seen as a distortion of the healthy speech signal. An often used distortion measure in speech enhancement is the Short Time Objective Intelligibility method (STOI), and its variant ESTOI (Taal et al., 2010). STOI is not directly applicable to pathological speech as STOI assumes that the distorted (here: pathological) signal and the reference signal have equal duration, which is seldom the case. Janbakhshi et al. (2019) proposed a modification of STOI and E-STOI, called P-STOI and P-ESTOI, which performs time alignment of the pathological and reference signals, and which can estimate severity with a high correlation to listener scores for two separate databases of dysarthric speech. Therefore, we include P-STOI and P-ESTOI in our comparison.

Recognising that advancements in speech enhancement evaluation can be applied to the evaluation of speech severity, we are also interested if we can apply techniques used in synthetic speech evaluation (i.e., naturalness evaluation) to oral cancer speech severity evaluation. Specifically, we investigate whether the most common objective approach used in synthetic speech evaluation, the Mel-cepstral distortion (MCD), can be used for the oral cancer speech severity estimation task (Kubichek, 1993).

Reference-free methods perform objective evaluation without the need for a transcription of the pathological speech signal or the need for a reference (healthy) speech signal (Woisard et al., 2022; Quintas et al., 2020; Bin et al., 2019; Zhou et al., 2012). Instead, they use a statistical model (e.g., a deep neural network or a LASSO model) and a feature representation to provide the severity estimate $\hat{x}$. We investigated the following possible features: (1) long-time average spectrum (LTAS), which has been used in the detection of pathological speech (Smith and Goberman, 2014; Master et al., 2006) and for the evaluation of the effect of speech therapy or surgery on the speech (Tanner et al., 2005). Moreover, in our previous studies, LTAS was successfully used to differentiate between oral cancer speech (Halpern et al., 2020) and healthy speech; and between dysarthric speech and healthy speech (Halpern et al., 2021). (2) Speaker embeddings, which have attracted a lot of attention recently (i-vector Martínez et al., 2013; Laaridh et al., 2017, 2018, x-vector Quintas et al., 2020, d-vector Wan et al., 2018), and seem to be useful for oral cancer speech intelligibility estimation (Quintas et al., 2020). (3) In our previous studies we found that naive listeners perceive high severity samples as having low naturalness (Halpern et al., 2021; Illa et al., 2021). Therefore, we investigate how reference-free synthetic speech evaluation methods that measure naturalness perform on the severity evaluation task, i.e., global variance (GV) (Toda and Tokuda, 2007) and modulation spectrum (MS) (Takamichi et al., 2014). We will compare each feature using a LASSO-based statistical model. The LASSO model is used to predict the severity measure $\hat{x}$ from the feature representation after training on the ground truth severity scores. We believe that using LASSO (1) allows for a fairer comparison of features than neural networks, where performance might be dependent on tuning, initialisation seeds, or the chosen network architecture, (2) and it is an explainable machine learning technique which is a common desire in clinical practice.

Both the reference-free and the reference-based approaches need large amounts of training data. Along with the reference transcriptions mentioned before, ASRs require large amounts of speech data, which are not available for all languages. Comparison-based approaches require a reference healthy speech signal. Reference-free approaches also require some form of human labelling, namely, the judgement of severity from the listeners. These resources are typically difficult to obtain. Therefore, it is important to consider whether we can reduce the cost of labelling. The secondary research question of the work is the following: **RQ2: Are other approaches available that require less labelled training data while giving a similar performance on the speech evaluation task?**.

We investigate two possible approaches: (1) Instead of predicting the severity directly, we could predict the probability of absence/presence (classification/detector task) of oral cancer speech. In our clinical experience, we find that lighter cases of oral cancer speech are difficult to tell apart from healthy speech. We hypothesise that

this would appear as a lower probability score during classification which could be correlated with the severity scores. This classification/detector task only needs binary labels, which are substantially easier and cheaper to acquire as no expert annotators are needed. In other words, we are interested in knowing whether detectors can achieve comparable performance to regressors. **(RQ2.1)**. (2) We propose to use severity ratings from naive listeners instead of expert listeners. There is a growing amount of evidence that crowdsourcing could be a cost-effective tool to collect data from non-expert listeners (Lansford et al., 2016; Lansford and Borrie, 2017; Carvalho et al., 2021). To that end, we investigate how far ratings from naive, non-expert listeners recruited through a crowdsourcing platform agree with those of expert listeners **(RQ2.2)**.

The paper is organised as follows. In Section 2, we define the terminologies concerning speech quality. In Section 3, we explain how we gathered the oral cancer dataset used in this research, and we perform an initial exploratory analysis on the reliability of the collected ratings. The section ends with a comparison of naive and expert listeners where we answer **RQ2.2**. Section 4 explains the experimental design to answer the research questions and includes a methodological summary for each technique. Finally, Section 5 presents and discusses the results from the perspective of each research question. The dataset in this paper and the evaluation recipes are publicly available.[1]

2. Terminological remarks

In the present study, we will use four terms to describe different aspects of the evaluated pathological speech: intelligibility, severity of the voice disorder, severity of the speech (severity), and naturalness.

Intelligibility: Following Duffy (2005), intelligibility is defined here as *the extent to which speech can be transcribed by (naive) listeners solely based on acoustic cues*. .

Severity of the voice disorder: Following American Speech Language Hearing Association (2023), a voice disorder is present when an individual expresses concern about having an abnormal voice that does not meet daily needs—even if others do not perceive it as different or deviant. The standard evaluation for the severity of the voice disorder is the GRBAS (Hirano, 1981) and the CAPE-V (Zraick et al., 2011). The GRBAS and the CAPE-V are standardised questionnaires, which ask to rate well-defined acoustic properties of the voice, such as roughness or breathiness. These questionnaires can only be completed by SLPs who receive training on the evaluation of these speech properties.

Severity of the speech disorder: The present study will focus on estimating the severity of the speech disorder (in short: severity). We follow the definition of Balaguer et al. (2019): severity is *the degree of the overall deterioration of the audible signal*. It is a measure aiming to combine both intelligibility and voice quality. Our usage of the term severity does not refer to the severity of the disease (e.g. the TNM stage of the tumour O'Sullivan and Shah, 2003) nor to the severity of the voice disorder.

Naturalness: Naturalness refers to the quality of computer-generated (synthesised) speech as defined by the International Telecommunications Union standard (Union, 1996). In other words, naturalness refers to the indistinguishability of human speech from computer speech. We think that naturalness is a closely related concept to the severity of the speech disorder based on the authors' previous studies, which showed that natural pathological speech received low naturalness scores (Illa et al., 2021; Huang et al., 2022). Therefore, we investigate the applicability of objective naturalness estimation techniques to objective severity estimation.

To summarise, severity of the speech disorder (severity for short) is the overall sense of "disordered" or "pathological" speech, and refers to disorders of voice and voicing specifically. A high severity of the speech disorder often correlates with a low "intelligibility" of speech and low "naturalness" of the speech.

To further clarify the distinctions between the terminologies, we interpret the definitions on a few examples:

- A pathological speaker with an extremely creaky or breathy voice quality whom is well understood by others would be classified as high intelligibility, high severity of the voice disorder, and high severity of the speech disorder.
- A nasalising pathological speaker has a very high level of intelligibility – as his/her speech remains understandable to the listener – but will be associated with an equally high level of speech severity because the speech signal will be strongly altered at the acoustic level (and perceptually). Given that the loudness and pitch of the nasalising speaker is not affected, the severity of the voice disorder is low.
- When the speaker is not understood by others, the speaker has a low intelligibility and high severity of both the voice and the speech disorder.

3. Dataset collection and analysis of the rating study

The following sections present the oral cancer database, its collection and the oral cancer speech severity rating by naive listeners obtained through crowd-sourcing and by speech-language pathologists (SLPs). This will be followed by an exploratory analysis of the collected ratings, which aims to investigate the reliability of the ratings. Moreover, we will answer research question (**RQ2.2**) whether the severity scores from naive listeners are comparable to those of speech-language pathologists.

3.1. Collection of the dataset

We manually collected 3 h of audio data containing English oral cancer speech from YouTube. The dataset includes 16 speakers. The presence of oral cancer speech was determined by the content of the video and the authors' (B.H., R.V.S., M.v.d.B.) clinical experience with such speakers. The audio was then manually cut to exclude music, healthy speakers, and non-American English speakers. All utterances were downsampled to 16 kHz, loudness normalised to −0.1 dB, and finally mixed from stereo to mono using the `sox` tool. Transcriptions were created manually starting from baseline transcriptions generated by the Baseline ASR system explained in Section 4.3.1.

We distinguish the utterances based on whether the annotator (B.H.) was able to transcribe the utterance (intelligible) or not (unintelligible). The unintelligible utterances will only be used for the reference-free techniques.

After preprocessing and splitting, the dataset contains a total of 840 transcribed 10-s (140 min) long utterances, and an additional 936 5-s long utterances (78 mins) of speech that are not transcribed. The dataset is partitioned into four different sets: a training and an evaluation set for both approaches (Reference-based and Reference-free). The reference-based evaluation set consists of the transcribed (intelligible) utterances while the reference-free evaluation set also includes untranscribed (unintelligible) utterances. The reference-free approaches are also evaluated on the reference-based evaluation, to compare all approaches once using the same test set. Please be reminded that all the utterances in the reference-based test set are individually rated and transcribed thus the discrepancy in the duration tested.

Table 1 provides the details of the training and evaluation sets such as the amount of audio and the number of utterances in each of the data sets. The selection of speakers in the reference-free evaluation was identical to the setup used in Halpern et al. (2020), except for the addition of three new speakers to the evaluation set. Identical to Halpern et al. (2020), for each speaker in the oral cancer dataset, matched (in terms of gender and number of audio recordings per

[1] https://karkirowle.github.io/oral_cancer_corpus/.

Table 1
Partitioning of the speakers into the training and evaluation set. RF stands for reference-free, and RB stands for reference-based. The **red** colour indicates female speakers, while the **blue** colour indicates male speakers. The column "Phonetic cover(age)" indicates the percentage of the different phonemes in the lexicon (CMUDict) that is present in the utterance by that speaker. The column "VoxCeleb control" contains the id of the control speaker from the VoxCeleb dataset, which is used only during the detection task. In the case of the reference-free models, scores are extrapolated (see Section 4.2.1) and trained with all available audio, therefore the number of rated utterances (parentheses) differs from the number of utterances used for training. Note that id006, id009 and id012 are speakers that are part of our dataset but not used in this paper. Best viewed in colour.

Speaker	Training RF	Training RB	Evaluation RF	Evaluation RB	Utterances included	Phonetic cover	VoxCeleb control
id001			✓	✓	10	79.49%	
id002	✓				8	Unintelligible	id10571
id003	✓	✓			8	82.05%	id10078
id004	✓				8	Unintelligible	id10111
id005			✓	✓	10	94.87%	
id007	✓	✓			8	87.18%	id11250
id008		✓	✓		8	92.31%	
id010			✓	✓	3	74.36%	
id011	✓	✓			8	92.31%	id10242
id013			✓		10	Unintelligible	
id014			✓	✓	2	87.18%	
id015			✓	✓	3	74.36%	
id016			✓	✓	10	84.62%	
id017			✓	✓	10	92.31%	
id018			✓	✓	10	71.79%	
id019			✓	✓	8	84.62%	
Total speakers (16)	5	4	11	9	–	–	–
Total utterances (124)	1632 (40)	636 (32)	84	66	–	–	–
Total audio used	2 h 16 min	1 h 46 min	54 min	7 min	–	–	–

speaker) controls from the VoxCeleb dataset were used (Nagrani et al., 2020). We chose VoxCeleb as the control because, similar to our oral cancer corpus, it was collected from YouTube. This allows exclusion of potential YouTube characteristics as a confounding factor in the healthy speech vs. oral cancer speech detection task. The selection of speakers in the reference-based evaluation follows the setup used in Halpern et al. (2022).

In this study, we have refrained from using k-fold validation because the speakers have wildly varying total recording durations, meaning that the results of the individual folds would depend too much on the presence of speakers with more audio recordings in the training dataset.

3.2. Selection of stimuli for questionnaire

In order to determine which approach works best for the oral cancer speech severity evaluation task, we need ground truth ratings. Because it would be too costly to get ratings for all oral cancer speech utterances, we selected a subset of the oral cancer speech utterances for rating by the naive listeners and the expert listener.

The subset of utterances for rating was created by adhering to the following:

1. (whenever possible), of the speakers in the evaluation set, 10 utterances will be rated;
2. (whenever possible), of the speakers in both training sets, 8 utterances will be rated
3. sentences are selected such that they cover the highest number of different phonemes for each speaker (*phonetic coverage*);
4. if there are multiple recordings available for a given speaker, at least one utterance from each recording is used to maximise channel variability for the speaker (*recording diversity*). It is important that a recording is the whole stretch of speech recorded at once, it is not the same thing as an utterance, i.e., a recording can have multiple utterances.

Please note that the recordings that do not have transcriptions, cannot be optimised for phonetic content. These recordings are manually cut without taking phoneme coverage into account. On the other hand, the recordings with transcriptions are optimised for phonetic content. In order to do so, first, all the words in the utterances were mapped to ARPABET phonemes (stress markers were ignored) using the CMUDict.[2]

In order to select for each speaker the set of utterances that has the largest phonetic coverage, a greedy algorithm is used to obtain an approximate solution in each step. The greedy algorithm selects the utterance which maximises the loss function:

$$\mathcal{L}(A, B, \text{new}) = |A \setminus B| + \alpha \cdot \mathbb{I}_{\text{new}},$$

where A is the set of phonemes in an utterance and B is the set of already covered phonemes. In other words, the difference of the number of elements (cardinality) in each set is calculated at each step to obtain the new phonemes. The parameter $\alpha \in \mathbb{R}^+$ is a hyperparameter, which can be tuned for each speaker separately. $\mathbb{I}_{\text{new}}$ is an indicator function, which takes on the value 1 if the recording is new, otherwise it is 0. This parameter controls the importance of new recordings over the importance of new phonemes: an $0 < \alpha \leq 1$ means: given an equal number of new phonemes in two different candidate utterances, the utterance coming from a new recording is preferred. Extending this logic, we can see that for any arbitrary α where α is $k \leq \alpha \leq k + 1$ ($k \in \mathbb{Z}^+$), the increase of α allows for losing k additional phoneme(s), if the selected recording is new in the other candidate utterance. For most speakers, we could obtain a selection that has all the recordings with $\alpha = 0.1$, in other words, there is no trade-off between the phonetic coverage and the recording diversity, with the exception of one speaker (id011), where we used $\alpha = 1.1$, this is equivalent of losing one phoneme.

The final selection for rating by the naive and expert listeners consisted of 98 intelligible utterances and 26 unintelligible utterances. These 124 utterances are henceforth referred to as "the stimuli".

3.3. Distribution of questionnaires

The rating study was administered via Qualtrics.[3] The 124 utterances were randomly assigned to one of five questionnaires (with one questionnaire containing only 24 utterances). We then distributed the five questionnaires through Prolific.[4] Prolific users could complete any number (i.e., between 1 and 5) of the questionnaires, but each only once. In total, we recruited 100 participants for each questionnaire, obtaining a total of 500 responses. The stimuli in each questionnaire were randomised for each participant in order to average out possible

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[3] https://www.qualtrics.com/.

[4] https://www.prolific.co/.

Table 2
Mean ($\bar{x}$) and standard deviation (s) of naive listener scores (top) and speech-language pathologist scores (bottom) obtained for each recording (multiple utterances are rated for each recording) and speaker rated in the rating study. Spk stands for the speaker id, Rec stands for the recording id.

Naive listener scores																
Spk	id001		id002						id003	id004				id005	id007	id008
Rec	1	3	8	12	14	16	19	25	10	11	15	27	29	18	21	23
$\bar{x}$	3.03	4.28	1.05	1.13	1.38	1.16	1.12	1.79	1.98	1.18	1.06	1.11	1.08	4.90	4.26	2.39
s	0.78	0.75	0.22	0.34	0.60	0.56	0.47	0.59	0.66	0.38	0.24	0.42	0.27	0.33	0.74	0.80
Spk	id008	id010	id011							id013	id014	id015	id016	id017	id018	id019
Rec	24	31	4	5	6	7	13	22	28	17	30	32	33	34	35	36
$\bar{x}$	2.60	4.21	4.06	4.08	4.66	4.21	4.31	3.73	3.59	1.29	4.57	3.28	3.84	4.75	2.44	3.90
s	0.75	0.73	0.76	0.72	0.53	0.78	0.74	0.95	0.83	0.48	0.57	1.09	0.86	0.67	0.71	0.75
Speech-language pathologist (SLP) scores																
Spk	id001		id002						id003	id004				id005	id007	id008
Rec	1	3	8	12	14	16	19	25	10	11	15	27	29	18	21	23
$\bar{x}$	3.5	4.5	1.0	1.0	1.5	1.0	1.0	1.0	2.24	1.0	1.0	1.0	1.0	5.0	3.91	2.33
s	0.5	0.5	0.0	0.0	0.5	0.0	0.0	0.0	0.65	0.0	0.0	0.0	0.0	0.0	0.68	0.47
Spk	id008	id010	id011							id013	id014	id015	id016	id017	id018	id019
Rec	24	31	4	5	6	7	13	22	28	17	30	32	33	34	35	36
$\bar{x}$	2.6	4.67	5.0	5.0	5.0	5.0	5.0	4.0	5.0	1.2	4.5	4.0	4.40	4.9	2.8	4.0
s	0.8	0.47	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.5	0.0	0.49	0.3	0.6	0.0

learning effects. We found only one American English expert SLP to rate the audio samples, which further emphasises the need for automatic evaluations. The SLP rated all 124 utterances using the same method as the naive listeners.

The task for each participant (both the naive and expert listeners) was to rate the severity (as defined in Section 2) of each utterance on a 5-point Likert scale. Participants received instructions before the study on how to rate the severity of the speech, which can be found in Appendix.

As the partial aim of the study was to investigate whether naive listeners could carry out evaluation of speech severity, we kept training of the listeners to a minimum, and only familiarised listeners with the task (and not the type of speech). Each questionnaire started with an example of a completely healthy utterance taken from the CMU Arctic corpus (Kominek and Black, 2004) (score 5), and an example of a particularly pathological utterance taken from the TORGO (Rudzicz et al., 2012) corpus of dysarthric speech (score 1). The speech severity of these utterances was discussed with an SLP to ensure that these examples were appropriate. Please note that due to the lack of publicly available oral cancer speech corpora and the need for ethical approval when using clinical data, we resorted to using a pathological speaker from the TORGO corpus. Another option would have been to exclude a speaker from our corpus. However, we did not want to exclude a speaker from our corpus as we already had a low number of speakers.

3.4. Results of the naive listener rating study

In this section, we carry out a preliminary analysis of the rating study. We start the analysis of the rating study results by assessing the consistency of the ratings, i.e., how similar are the raters to each other. Then, we investigate whether we can observe any global tendency of the ratings, e.g., whether the raters had a tendency to choose the extremities of the scales. Finally, we look at the consistency of the ratings on the speaker-level, i.e., for different recordings of the same speaker, did the raters give different scores?

For investigating the consistency or agreement in naive listeners' ratings we first calculated the interrater correlation (IRR) by averaging all pairwise Pearson's correlations for all pairs of raters. As the raters might be overlapping between the questionnaires, we calculate the IRR for each questionnaire separately. A high IRR means a high level of agreement in the ratings. Second, we carried out a principal component analysis (PCA) to visualise any clusters of raters who might have followed similar rating strategies. For example, if we were to find two visually separable clusters, we could conclude that the raters rated the samples according to two different interpretations of the rating task. Fig. 3 shows the interrater correlations and the results of the principal components analysis. The IRRs vary between 0.83 and 0.87, indicating good consistency between ratings. Furthermore, the PCA does not show multiple clusters of ratings, meaning that the raters did not follow substantially different rating strategies.

A commonly heard criticism of uneven Likert scales is that the "neutral" (3 in our case) score is used as a fall-back option and hence has the highest frequency. To assess if there is any global tendency within the results, we looked at the mean scores for each recording in Table 2 and the histogram of the dataset, see Fig. 1. The lowest mean score was for recording 8 (1.05), while the highest was for recording 18 (4.90), which indicates that participants used the full extent of the rating scale. Furthermore, the histogram in Fig. 1 shows that a rating of 5 (healthy speech) was most commonly used. It is true that the obtained ratings do not seem to exhibit a completely uniform distribution, but this is more likely due to the fact that the severity of the utterances was not controlled when selecting the utterances.

We then carried out an analysis of the range of the means of the ratings of the recordings per speaker. The upper part of Table 2 lists all the mean scores for each recording, grouped by speakers. There are 5 speakers (id001, id002, id004, id008, id011) who have multiple recordings. Three of them have a score range of more than 0.5: id001 (range = 1.25), id011 (range = 1.07), id002 (range = 0.74). In the case of id002, there are two recordings (Recording 14 and 25) that seem to receive noticeably higher scores (1.79 and 1.38) than the other recordings of the speaker.

A Wilcoxon signed-rank hypothesis test (see Table 3 for the p-values) showed a significant difference between the distribution of scores for most speakers except id004. In the case of id002, most recordings were consistent, except recordings 14 and 25.

There might be multiple reasons why there were inconsistencies for different recordings, however, the scores seem to be well aligned with the scores of the expert listener (see Section 3.5), therefore, we think these differences reflect actual differences in speech severity, and that some of these differences might be explained by differences in time of the recordings rather than inconsistencies in the ratings by the naive listeners. For example, speakers id011 and id001 self-report that their videos were recorded at different moments in time, where they have a different speech severity, which might explain the rather large range for these two speakers. In the case of id002, informal listening by author B.H. confirmed that recordings 14 and 25 indeed

Table 3
Adjusted p-values of the Wilcoxon signed-rank between the naive score distributions of the recordings. Adjustment was done using the Dunn–Sidak correction with $\alpha = 0.05, \alpha_{adj} = 0.00024$. We denote $p_{adj} < \alpha$ as *. Highlighted cells indicate score distributions for recordings of the same speaker. Best viewed in colour.

	id001		id002						id004				id008		id011						
	1	3	8	12	14	16	19	25	11	15	27	29	23	24	4	5	6	7	13	22	28
1	Same	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
3	*	Same	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	1.0	1.0	*	*
8	*	*	Same	0.96	*	1.0	1.0	*	0.063	1.0	1.0	1.0	*	*	*	*	*	*	*	*	*
12	*	*	0.96	Same	*	1.0	1.0	*	1.0	1.0	1.0	1.0	*	*	*	*	*	*	*	*	*
14	*	*	*	*	Same	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
16	*	*	1.0	1.0	*	Same	1.0	*	1.0	1.0	1.0	1.0	*	*	*	*	*	*	*	*	*
19	*	*	1.0	1.0	*	1.0	Same	*	1.0	1.0	1.0	1.0	*	*	*	*	*	*	*	*	*
25	*	*	*	*	*	*	*	Same	*	*	*	*	*	*	*	*	*	*	*	*	*
11	*	*	0.06	1.0	*	1.0	1.0	*	Same	0.11	1.0	0.32	*	*	*	*	*	*	*	*	*
15	*	*	1.0	1.0	*	1.0	1.0	*	0.11	Same	1.0	1.0	*	*	*	*	*	*	*	*	*
27	*	*	1.0	1.0	*	1.0	1.0	*	1.0	1.0	Same	1.0	*	*	*	*	*	*	*	*	*
29	*	*	1.0	1.0	*	1.0	1.0	*	0.32	1.0	1.0	Same	*	*	*	*	*	*	*	*	*
23	*	*	*	*	*	*	*	*	*	*	*	*	Same	*	*	*	*	*	*	*	*
24	*	*	*	*	*	*	*	*	*	*	*	*	*	Same	*	*	*	*	*	*	*
4	*	*	*	*	*	*	*	*	*	*	*	*	*	*	Same	1.0	*	*	*	*	*
5	*	*	*	*	*	*	*	*	*	*	*	*	*	*	1.0	Same	*	*	*	*	*
6	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	Same	*	*	*	*
7	*	1.0	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	Same	1.0	*	*
13	*	1.0	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	1.0	Same	*	*
22	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	Same	0.20
28	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	0.20	Same

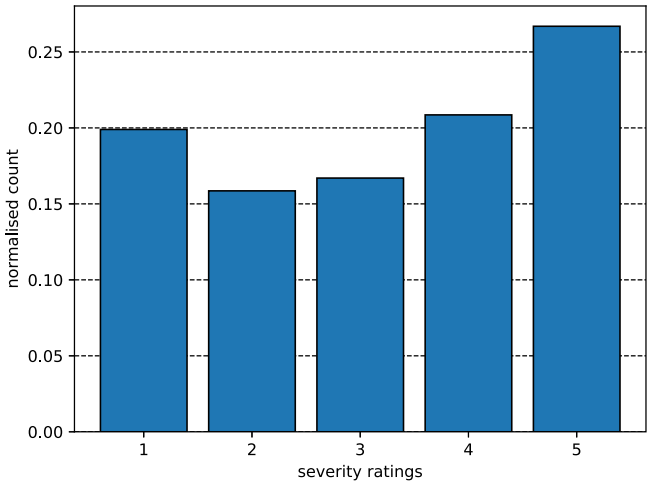

Fig. 1. Histogram of all the ratings in our dataset. The x-axis shows the ratings given, 1 being the most severe, 5 being the least severe or healthy, the y-axis shows the normalised counts.

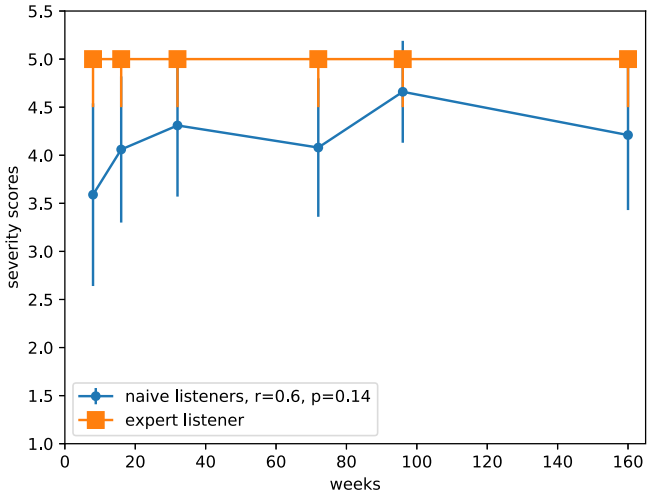

Fig. 2. Temporal evolution of the ratings of id011 according to the naive listeners (blue, with circular markers) and the expert listeners (orange, with square markers). One out of the seven recordings of id011 did not have temporal metadata, which was omitted from this analysis.

were a lot more intelligible than the other recordings from speaker id002. We hypothesise that this is because the recordings were done at different times, however, this is not obvious from the content nor the corresponding metadata of the recordings. As we had metadata for six recordings of id011, we decided to quantify the Pearson's correlation between the number of weeks since the surgery and the obtained ratings, and visualised the temporal evolution of these ratings in Fig. 2. We found that in the case of naive listeners, there was a moderate but insignificant ($r = 0.51, p = 0.3$) correlation between the weeks and the severity score, while the expert listener gave a rating of 5 to all recordings. Taken together, there is no correlation between the weeks and the ratings.

3.5. RQ2.2: Comparison of naive and expert listeners

In order to compare the naive and expert listener scores, we used a Pearson's correlation between the mean of the scores from the naive listeners and the mean of the scores from the SLP. The strength of the correlation was $r = 0.92$ ($p < 0.001$), which is very high. This strongly indicates that we can use ratings from naive listeners obtained through crowdsourcing to rate the severity of the speech reliably on a 5-point scale. For the rest of this paper, we will use the naive listener scores as ground truth severity scores to validate our different methods, as these scores are based on more raters and are thus more granular than that of a single SLP.

Please note, in general, we expect that differences between the ratings of the naive listener and SLP would be much more apparent in evaluation questionnaires that ask for explicit voice qualities such as breathiness, (i.e., as in GRBAS Oates, 2009), as naive listeners have little understanding about the acoustic cues corresponding to these terminologies.

4. Methods

4.1. Experimental design

Table 4 lists all models that were compared in order to find the best technique for oral cancer speech severity evaluation. For each model,

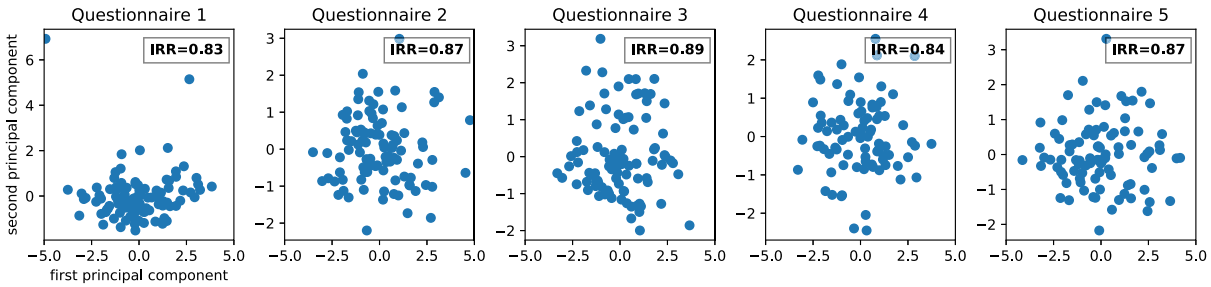

Fig. 3. Principal component analysis of the raters for each questionnaire. Note that the y-axis is different on the first plot to show certain outliers. The IRR in the upper right corner stands for the intrarater correlation.

Table 4
Overview of all systems evaluated in this paper. ✗ means reference-free, ✓ means reference-based. Reference-free models are trained with the reference-free dataset, and reference-based models are trained with the reference-based dataset.

Model	Reference	Reference type
GV-detector	✗	No
GV-regressor	✗	No
MS-detector	✗	No
MS-regressor	✗	No
LTAS-detector	✗	No
LTAS-regressor	✗	No
xvec-detector	✗	No
xvec-regressor	✗	No
dvec-detector	✗	No
dvec-regressor	✗	No
Baseline	✓	Transcription
Baseline+OC	✓	Transcription
DNN for AM Retraining	✓	Transcription
fMLLR	✓	Transcription
MCD	✓	Synthetic
P-STOI	✓	Synthetic
P-ESTOI	✓	Synthetic

we indicate whether it uses a reference (see column "Reference"), and if so, what type of reference (column "Reference type"). For the ASR reference-based experiments (Baseline, Baseline+OC, DNN for AM Retraining, fMLLR), there are additional variants that we have not listed in the table for the sake of clarity, please see Section 4.3.1 for more details.

All models will be compared on the reference-based evaluation set while the reference-free models will also be compared on the reference-free evaluation set. In order to find the best technique for oral cancer speech severity evaluation, the severity estimate $\hat{x}$ obtained for each model is correlated with the average severity rating obtained from the naive listeners.

Note that data augmentation could be potentially used to improve the performance of some models, however, this would have prohibited a fair comparison of approaches in our study. For example, the MCD would hardly benefit from the data augmentation as it does not use any training data. Therefore, we consider such modifications out of scope for the present study.

4.2. Reference-free approaches

To evaluate the reference-free approaches in a consistent way, we will use a LASSO-based detection and regression model (Section 4.2.1). The LASSO model will be tested with the d-vector (dvec), x-vector (xvec) (Section 4.2.2), LTAS (Section 4.2.3), and the global variance and the modulation spectrum (Section 4.2.4) features.

4.2.1. Reference-free experimental setup

LASSO is a variant of linear regression (Tibshirani, 1996), which performs feature selection and regression simultaneously. Potentially, for a given linear regression task, some features do not contain any relevant information to make predictions or contain information that is collinear with the other features, causing overfitting. In LASSO, coefficients of regression are encouraged to be close to zero if they do not provide useful information. Zeroing (pruning) some features means that the model requires only a subset of all predictors, making the statistical model parsimonious and easier to interpret.

There are two variants of LASSO that we will use, one for regression and one for detection. For regression, we will use the vanilla LASSO. At inference time, vanilla LASSO's computation is identical to linear regression (Eq. (1)), however, at training time the coefficients ($\mathbf{w} \in \mathbb{R}^m$ where m is the dimensionality of the feature) are obtained in a slightly different way by adding the sparsity penalty to the ordinary least squares loss function (see Eq. (2)):

$$\hat{x}_i = \mathbf{w}^T \mathbf{h_p}(i), \tag{1}$$

$$\hat{\mathbf{w}} = \text{argmin}_{\mathbf{w}} \|\mathbf{w}^T \mathbf{h_p}(i) - x\|_2^2 + \lambda \|\mathbf{w}\|_1. \tag{2}$$

Pruning of the features is facilitated by setting the parameter $\lambda = 0.1$: the larger this parameter is, the closer the coefficients are to zero.

For detection, we will use a logistic LASSO, which is similar to LASSO with two key differences: (1) the addition of the sigmoid function to obtain the detection probability; (2) instead of x, binary labels are used, which we denote with $x_b \in \{0, 1\}$. Chunks in the VoxCeleb dataset take on $x_b = 1$, while chunks in the oral cancer corpus take on $x_b = 0$.

$$\hat{x}_i = \sigma(\mathbf{w}^T \mathbf{h_p}(i)) \quad \sigma(x) = \frac{1}{1 + \exp(-x)}, \tag{3}$$

$$\hat{\mathbf{w}} = \text{argmin}_{\mathbf{w}} \|\sigma(\mathbf{w}^T \mathbf{h_p}(i)) - x_b\|_2^2 + \lambda \|\mathbf{w}\|_1. \tag{4}$$

Note that this model is effectively a perceptron with L_1 regularisation. The addition of the sigmoid function does not cause any problems with the optimisation, as the sigmoid function is differentiable with respect to $\mathbf{w}$.

In order to compute the different reference-free features, we first chunk the utterances into 5 s segments ($\mathbf{y}_p(i), i \in [1, T]$, where i is the chunk index, and T is the total number of chunks) – note that the last chunk duration can be shorter than 5 s. Subsequently, different features are extracted (see the later sections in Section 4.2) for each of these 5-s chunks, where we obtain $\mathbf{h}_p(i) \in \mathbb{R}^d$, where d depends on the kind of feature we are using, and i denotes the chunk index. Therefore a training pair consists of the chunk of the recording $\mathbf{h}_p(i)$, and the corresponding severity score x.

The final prediction scores are obtained on the level of the recordings. Using recordings instead of utterances is more sound from a clinical linguistic perspective, as this approach takes into account the fact that the impact of oral cancer surgery on pronunciation is different for different sounds. Therefore, the perceived severity should be also different for the different parts of the recordings as they contain different sounds.

To reflect this, we create a recording-level score $\hat{x}$ from all the available utterance chunks within a recording. The final score is obtained

as the weighted average of the scores for each utterance, i.e., for a recording that has n utterances with a number of chunks T_1, T_2 and T_n:

$$\hat{x} = \frac{1}{T_1 + T_2 + \cdots + T_n}\left(\sum_{i=1}^{T_1}\hat{x}_{1,i} + \sum_{j=1}^{T_2}\hat{x}_{2,j} + \cdots + \sum_{k=1}^{T_n}\hat{x}_{n,k}\right). \tag{5}$$

Each reference-free LASSO model was trained with the Reference-free training set, which includes both the intelligible and the unintelligible utterances. As only a selection of the utterances in the corpus, and thus of the reference-free training set, was rated by human listeners, we extrapolated these ratings for those utterances without ratings in order to increase our training set size. All utterances without a rating received the average rating calculated over all rated utterances of the same recording of that speaker. The extrapolated ratings were also used as ground truth ratings, and are referred to as x.

4.2.2. Speaker embeddings

The two speaker embedding features tested in this work are the angular x-vector (xvec) (which is an improved version of the x-vector) and the d-vector (dvec) (Bredin et al., 2020; Coria et al., 2020). To extract a speaker embedding, the $\mathbf{y}_p(i)$ was fed through a deep neural network (DNN). Instead of the class labels, the activations of one of the intermediate layers were extracted and used as the speaker embeddings feature $\mathbf{h}_p(i)$ in our LASSO model.

Angular x-vectors differ from the conventional x-vector model (Snyder et al., 2018) by using an angular softmax function instead of the normal softmax function, and using SincNet features instead of MFCCs (Ravanelli and Bengio, 2018). The d-vector uses the generalised end-to-end loss (GE2E) as its loss function, while having 40-dimensional Mel filterbank features. For both of these models, we used (previously) publicly available implementations[5]

4.2.3. LTAS

In order to obtain the LTAS features, we extracted a (so-called) Kaldi spectrogram from the audio chunk $\mathbf{y_p}(i)$ with a 25 ms length Povey window, 10 ms frame shift and 256 frequency bins using the PyTorch torchaudio library. The obtained spectrogram is denoted by $\mathbf{S}_p \in \mathbb{R}^{256\times L}$, where 256 is the number of frequency bins and L is the number of analysis frames in the spectrogram, which is dependent on the duration of the individual chunks. We obtain the LTAS vector by stacking the mean and standard deviation for all 256 frequency bins which results in a $\mathbf{h}_p \in \mathbb{R}^{512}$ LTAS vector:

$$\mathbf{h}_p = \begin{bmatrix} \frac{1}{L}\sum_{j=0}^{L-1}\mathbf{S}_p(0,j) \\ \vdots \\ \frac{1}{L}\sum_{j=0}^{L-1}\mathbf{S}_p(255,j) \\ \sqrt{\frac{\sum_{j=0}^{L-1}\mathbf{S}_p(0,j)-\mathbf{h}_p(0)}{L-1}} \\ \vdots \\ \sqrt{\frac{\sum_{j=0}^{L-1}\mathbf{S}_p(255,j)-\mathbf{h}_p(255)}{L-1}} \end{bmatrix} \tag{6}$$

4.2.4. Global variance and modulation spectrum

Both the global variance (GV) and modulation spectrum (MS) are commonly used to evaluate synthetic speech objectively. For the GV calculation, we first calculated 20-dimensional librosa MFCC trajectories ($\mathbf{c}_p(i) \in \mathbb{R}^{20\times M}$) from the audio chunks $\mathbf{y}_p(i)$. From each MFCC trajectory, we calculated a time-axis variance estimate, which resulted in the 20-dimensional GV features, using:

$$\mathbf{h}_p(i) = \frac{1}{M}\sum_{j=0}^{M-1}\mathbf{c}_p(i)(j) - \bar{\mathbf{c}_p}(i) \quad \bar{\mathbf{c}_p}(i) = \sum_{j=0}^{M-1}\mathbf{c}_p(i)(j). \tag{7}$$

For the MS, we used the implementation from nnmnkwii (Yamamoto et al., 2021). First, we extracted 60-dimensional Mel-generalised cepstrum coefficients (MGC), and ignored the 0th order MGC. Subsequently, we took the power of the discrete Fourier transform of the MGC parameter trajectory across the time-axis. To obtain a duration-independent feature, we took the time-axis average, which resulted in the final 59-dimensional MS feature, which was computed using:

$$\mathbf{h}_p(i) = \frac{1}{M}\sum_{i=0}^{M-1}(\mathcal{F}\{\mathbf{c}_p\})^2(i). \tag{8}$$

4.3. Reference-based

4.3.1. Word-level ASR systems

We used four different ASR systems from our previous work (Nagrani et al., 2021) to generate word-level transcriptions for oral cancer speech recordings: (1) Baseline; (2) Baseline + oral cancer (Baseline + OC); (3) DNN for AM retraining; and (4) feature-space maximum likelihood linear regression (fMLLR) based system (see Section 4.3.1). All four ASR systems were trained by leveraging data from Wall Street Journal (WSJ) speech corpus (healthy speech) and oral cancer speech except the baseline system, which was trained only on WSJ speech. We wanted to use a spontaneous corpus for pretraining, however we were only aware of the Switchboard corpus, which consists of telephone conversations between speakers of American English. However this dataset has a lower sampling rate (8 kHz) than our audio (16 kHz), which would have likely affected the results. For example, in Halpern et al. (2020), we found that sibilants are important indicators of oral cancer speech, which often have acoustic cues in the high frequency range.

For each system, we created a variant with a tri-gram language model and an RNN (LM). The Levenshtein distance has previously been found to perform well for speech severity evaluation using ASR systems (Tripathi et al., 2020). Therefore, here we used this same measure. The Levenshtein distance was calculated between the ground truth transcription and decoded transcription of each utterance, and subsequently correlated with the average rating from the naive listeners.

Baseline: The baseline system is a standard hybrid DNN-HMM ASR system which is trained exclusively on healthy speech using the `si284` set of the WSJ corpus (Paul and Baker, 1992). The acoustic model (AM) of the baseline system consisted of 5 feed-forward hidden layers of dimension 1500 and a softmax output layer of 3431 (equal to the number of HMM states). The input features to the DNN AM were 23 dimensional filterbank plus 3 dimensional pitch features (FB+P). The 3 dimensional pitch feature consists of a probability of voicing, pitch and delta-pitch feature (Ghahremani et al., 2014). We followed the Kaldi recipe[6] in training the baseline DNN AM.

Baseline + OC: The system baseline + OC followed a similar training pipeline as the baseline system, with the exception of using both the WSJ `si284` data and the oral cancer training data to train the DNN AM.

DNN for AM retraining: The DNN for AM retraining system was based on the baseline system (in Section 4.3.1) and sequently retrained. Specifically, the baseline DNN-HMM AM was used to generate forced-alignments for the oral cancer training data using its reference transcriptions. Next, the oral cancer training speech and its corresponding alignments were taken as training data and labels to re-train the DNN-HMM AM.

fMLLR: The fMLLR system aimed at leveraging the success of the fMLLR algorithm in speaker adapted feature (named fMLLR feature) generation (Gales, 1998). In the context of oral cancer speech recognition, the use of fMLLR features could suppress pathological

[5] https://huggingface.co/hbredin/SpeakerEmbedding-XVectorMFCC-VoxCeleb. https://github.com/resemble-ai/Resemblyzer.

[6] `kaldi/egs/wsj/s5`.

speech sound characteristics in oral cancer speech, encouraging oral cancer speech representations to be more similar to those of normal speech, hence improving the recognition performance (Halpern et al., 2022). Similar to the baseline + OC system, the fMLLR system was trained using both the WSJ and oral cancer speech data, with the only difference being of applying fMLLR features (40 dimensions) instead of FB+P features during DNN AM training. The fMLLR features were estimated during GMM-HMM training, also using the merged WSJ and oral cancer speech data. The initial inputs of the fMLLR features were 39-dimensional MFCC+$\Delta + \Delta\Delta$ features.

4.3.2. Comparison-based approaches

The comparison-based approaches require a reference, healthy speech signal ($\mathbf{y}_r \in \mathbb{R}^{d_r}$, where d is the duration of the reference signal) along with the pathological signal ($\mathbf{y}_p \in \mathbb{R}^{d_p}$ where d_p is the duration of the pathological signal). Because there are no healthy references available, synthetic healthy references are generated using a highly natural Tacotron-2 text-to-speech synthesis (TTS) system.[7] (Shen et al., 2018)

P-STOI and P-ESTOI: Both P-STOI and P-ESTOI are modifications of the STOI technique, commonly used in the speech enhancement field (Taal et al., 2010). STOI does not account for the different tempi of healthy and pathological speech and assumes time-aligned speech signals. To account for the time-alignment issue, P-STOI and P-ESTOI extend the STOI technique with dynamic time warping (DTW). The calculation of the P-STOI/P-ESTOI scores is as follows. First, we extract the 1/3 octave band time-frequency (TF) representation $\mathbf{H}_p$ and $\mathbf{H}_r$ from $\mathbf{y}_r$ and $\mathbf{y}_d$, where we align $\mathbf{H}_r$ and $\mathbf{H}_p$ using DTW. We estimate the cross-correlation between the aligned representations. As these representations are two-dimensional, the cross-correlation can be done along either the temporal or the spectral axis. The temporal estimate is called the P-STOI score, while the spectral estimate is called the P-ESTOI score (Janbakhshi et al., 2019). The estimated scores are used as our severity measure $\hat{x}$.

MCD: The Mel-cepstral distortion (MCD) metric is usually used to measure the difference between a synthetic and a natural speech signal in order to objectively evaluate synthesis quality in TTS development. Here, the MCD metric is used to measure the difference between the pathological speech signal $\mathbf{y_p}$ and the reference speech signal $\mathbf{y_r}$ in order to predict the severity score $\hat{x}$. To calculate the MCD, we first extracted 20-dimensional Mel frequency cepstral coefficients (MFCCs) from $\mathbf{x_p}$ and $\mathbf{x_r}$ using the librosa Python library (McFee et al., 2020). We denote the obtained representations with $\mathbf{H_p} \in \mathbb{R}^{20\times M}$ and $\mathbf{H_r} \in \mathbb{R}^{20\times L}$ where L and M represent the number of analysis frames in the MFCC. The reference and pathological MFCCs have to be aligned if they have different lengths, otherwise calculation of the MCD is impossible. Therefore dynamic time warping (DTW) is performed to align the MFCCs. The aligned reference MFCC is denoted as $\mathbf{H_{rp}} \in \mathbb{R}^{20\times M}$. Following standard procedure, the α scaling coefficient was used (Mashimo et al., 2001). Note that the zeroth-order MFCC is ignored following standard practice because it is dependent on the gain of the speech, which can be sensitive to noise.

$$\hat{x} = \mathrm{MCD}(\mathbf{H_p}, \mathbf{H_{rp}}) = \frac{\alpha}{M} \sum_{i=0}^{M-1} \sqrt{\sum_{j=1}^{19} (\mathbf{H_p}(i,j) - \mathbf{H_{rp}}(i,j))^2} \quad \alpha = \frac{10\sqrt{2}}{\ln 2} \tag{9}$$

5. Results

5.1. RQ1: Comparison of all approaches on the speech severity evaluation task

Table 5 lists the Pearson's correlations of the estimated severity score of all approaches with the average human rating of the naive listeners. All results are obtained on the reference-based evaluation set. The table is divided into three blocks. The upper part of the table shows the reference-free, the lower part of the table shows the reference-based approaches in two blocks: one block is for the ASR models, the other block includes the comparison-based approaches. When a model has a higher Pearson's correlation than another model, we will say that it outperforms the other model.

Comparing all approaches, we see that the DNN for AM retraining models performed the best on the reference-based evaluation set. This

[7] https://github.com/NVIDIA/tacotron2.

Table 5
Pearson's correlation of all the approaches evaluated on the reference-based evaluation set, rounded to two decimals. A **cyan** background colour marks the ASR acoustic models which use oral cancer data during training. "TTS reference" indicates whether a synthetic speech ground truth is used. The best performing model is emphasised with a **bold** typeface. Best viewed in colour.

Model	Pearson's r	p	Language model	TTS reference
Reference-free approaches				
LTAS-detector	−0.25	***	N/A	N/A
LTAS-regressor	0.52	***	N/A	N/A
dvec-detector	0.55	***	N/A	N/A
dvec-regressor	0.47	***	N/A	N/A
xvec-detector	0.13	0.02	N/A	N/A
xvec-regressor	0.11	0.057	N/A	N/A
GV-detector	0.28	***	N/A	N/A
GV-regressor	0.28	***	N/A	N/A
MS-detector	0.29	***	N/A	N/A
MS-regressor	0.64	***	N/A	N/A
Reference-based approaches (ASR-based)				
Baseline	0.72	0.02	n-gram	Ø
Baseline + OC	0.75	0.005	n-gram	Ø
DNN for AM Retraining	**0.80**	0.006	n-gram	Ø
fMLLR	0.62	0.06	n-gram	Ø
Baseline	0.68	0.03	RNN	Ø
Baseline + OC	0.67	0.03	RNN	Ø
DNN for AM retraining	0.72	0.02	RNN	Ø
fMLLR	0.49	0.15	RNN	Ø
Reference-based approaches (comparison-based)				
MCD	0.27	0.45	N/A	Yes
P-STOI	−0.25	0.49	N/A	Yes
P-ESTOI	0.13	0.72	N/A	Yes

***Indicates p-value $< 10^{-3}$, otherwise p-value is provided.

Table 6
Pearson's correlation of the reference-free approaches on both the RB and RF evaluation sets. Of each detector/regressor pair, **red** background indicates a worse correlation while **green** indicates a better correlation than the other member of the pair. The best performing model is emphasised with a **bold** typeface for each evaluation type (reference-free and reference-based). Note that the data in the right column of the table is identical to the top part of Table 5, we present the data twice for ease of understanding. Best viewed in colour.

Model	Reference-free evaluation		Reference-based evaluation	
	Pearson's r	p	Pearson's r	p
Reference-free approaches				
GV-detector	0.64	***	0.28	***
GV-regressor	0.72	***	0.28	***
MS-detector	0.68	***	0.29	***
MS-regressor	**0.76**	***	**0.64**	***
LTAS-detector	0.27	***	−0.25	***
LTAS-regressor	0.66	***	0.52	***
dvec-detector	0.46	***	0.55	***
dvec-regressor	0.70	***	0.47	***
xvec-detector	0.55	***	0.13	0.02
xvec-regressor	0.53	***	0.11	0.057

***Indicates p-values $< 10^{-3}$, otherwise p-value is written.

means that reference-based approaches seem to outperform reference-free approaches in determining oral cancer speech severity when a reference is available for evaluating the speech severity. We will further discuss the possible reasons for the performance differences of the ASR models in Sections 5.3 (data differences), and 5.4 (language model differences).

The reference-free approaches achieved low to moderate correlations with the average listener scores on the reference-based evaluation set: the best approach was the MS-regressor, followed by the dvec-detector and the LTAS-regressor. We did not observe any obvious patterns in these results, so these will not be further discussed. Finally, most comparison-based approaches performed quite poorly on the reference-based evaluation. We will discuss these results in Section 5.5.

Table 6 shows the results for the reference-free detector and regressor approaches on the reference-free evaluation set. The left column shows that the best approach on the reference-free evaluation set was the MS-regressor, followed by the GV-regressor and the dvec-regressor. (**RQ1**). For both the regression and the detection task, the best features are those that are used in the evaluation of synthetic speech. We will further discuss the general implications of this in Section 6.

5.2. RQ2.1: Can detectors achieve comparable performance to regressors on the speech severity evaluation task?

Comparing the correlations of the regressors with the detectors on the reference-based evaluation set (right columns of Table 6) and reference-free evaluation set (left columns of Table 6), we observe that the regressors generally achieved higher correlations than the detectors, with the following exceptions: the xvec on the (both evaluation), dvec detector (reference-based evaluation), GV (reference-based evaluation). Overall (combining the reference-free and reference-based evaluation) the regression experiment was better in 60% of the cases.

These results show that the regressor models which were trained on the severity scores rather than the binary scores as the detectors were, are more informative for and better at the oral cancer severity evaluation task. Therefore, using binary class labels instead of severity scores is not a good solution when one wants to build automatic methods to evaluate the severity of oral cancer speech that have a good correlation with human ratings of the severity of the oral cancer speech.

5.3. Oral cancer data seems to help in ASR-based oral cancer severity evaluation

From Table 5 we can see that the model that has the highest correlation with the human ratings is a model that uses oral cancer data during training of the acoustic models (DNN AM Retraining), irrespective of the type of language model used. We expect that adding some oral cancer data to the training material is beneficial to the ASR models because the acoustic models then capture some of the mild disfluencies due to oral cancer speech in a vein that is similar to how human listeners quickly adapt to mild disfluencies in healthy speech (Kim and Nanney, 2014; Kim, 2015). It is interesting to note that even though the fMLLR uses oral cancer speech, it always achieves worse performance than the Baseline. We hypothesise that fMLLR adapts to the severity of the speaker and as such is able to learn the deviant pronunciations of an oral cancer speaker. Since fMLLR takes into account the entire recording of the speaker, this may result in an "overadaptation" to the oral cancer speaker. On the other hand, human listeners only hear a single utterance of a speaker at any given time, which is not enough to adapt to the deviant pronunciations of the oral cancer speaker. Consequently, the scores provided by the fMLLR models do not correlate that well with the human ratings compared to the models without fMLLR.

5.4. Weaker language models seem to lead to improved correlations

It is often implicitly assumed that language modelling will affect the intelligibility estimation, however, we are only aware of the study of Maier et al. (2009) which reports that some language modelling is beneficial for severity estimation, but the performance saturates after $n = 2$ with an n-gram. In our case, we can see that n-gram based language models outperformed the RNN-based language models in all cases. A more complex language model thus does not generally improve the correlation with listener scores. This makes sense: a stronger language model (here the RNN) will help the ASR to decode the acoustic signal using stronger lexical and semantic information than a weaker LM. This means that a model that uses a stronger LM will rely less on acoustic cues, while these acoustic cues are more helpful for severity evaluation. The results are consistent with the result of Maier et al. (2009) mentioned above. Therefore, we think that only language models with low complexity should be considered in speech severity evaluation, if considered at all.

5.5. Comparison-based methods seem to be lacking in performance

In general, we can see that the comparison-based approaches, i.e., MCD, P-STOI, and P-ESTOI, performed poorly compared to the other approaches. We hypothesise that this might be due to the DTW, which is used in all of the comparison-based techniques. We think that the DTW might not be a robust aligner in the noisy conditions that are sometimes present in the dataset. In the case of the comparison-based approaches, we align TTS speech (trained on read speech) with pathological speech (spontaneous speech). However, in the original P-STOI method, read speech was aligned with read speech. Potentially, the alignment is more difficult when the speech types do not match. Therefore, future work should look at other alignment methods (such as attention), and use multiple references to test their robustness under noisy conditions. It is likely that at least the P-STOI and P-ESTOI would improve when using multiple references as these methods are normally used with more references, however, that would have been an unfair comparison in the current study.

6. Discussion

In this paper, we built and compared multiple automatic speech evaluation systems for the evaluation of the severity of a speech impairment due to oral cancer, based on spontaneous speech.

Our main research question concerned finding the best method for the automatic evaluation of oral cancer speech. The best method for the automatic evaluation of oral cancer speech was the modulation spectrum regressor, for which no reference transcription is needed. If reference transcriptions are available, then automatic speech recognisers can be used, which showed the highest correlation with the naive listener ratings on the reference-based evaluation.

The majority of the methods showed a moderate to high correlation with the naive listener ratings, depending on the type of evaluation (reference-based, reference-free) used. These scores are, however, considerably lower than one would normally find for clean, read pathological speech (Janbakhshi et al., 2019; Windrich et al., 2008; Quintas et al., 2020). Our less good results are most likely due to the spontaneous nature of the speech used in our experiments and the fact that the recordings were obtained from YouTube and contained substantial background noise. Therefore, for a clinical use case, we would recommend using these models in tandem with the more predictive read-speech approaches (such as Quintas et al. (2020)) to obtain a more complete picture of the patients' speech difficulties.

While intelligibility is a factor in severity of the speech disorder (and also part of our questionnaire), the magnitude of the observed correlations indicates that there might be something more going on, i.e., the ASR systems might also be able to capture some aspects of

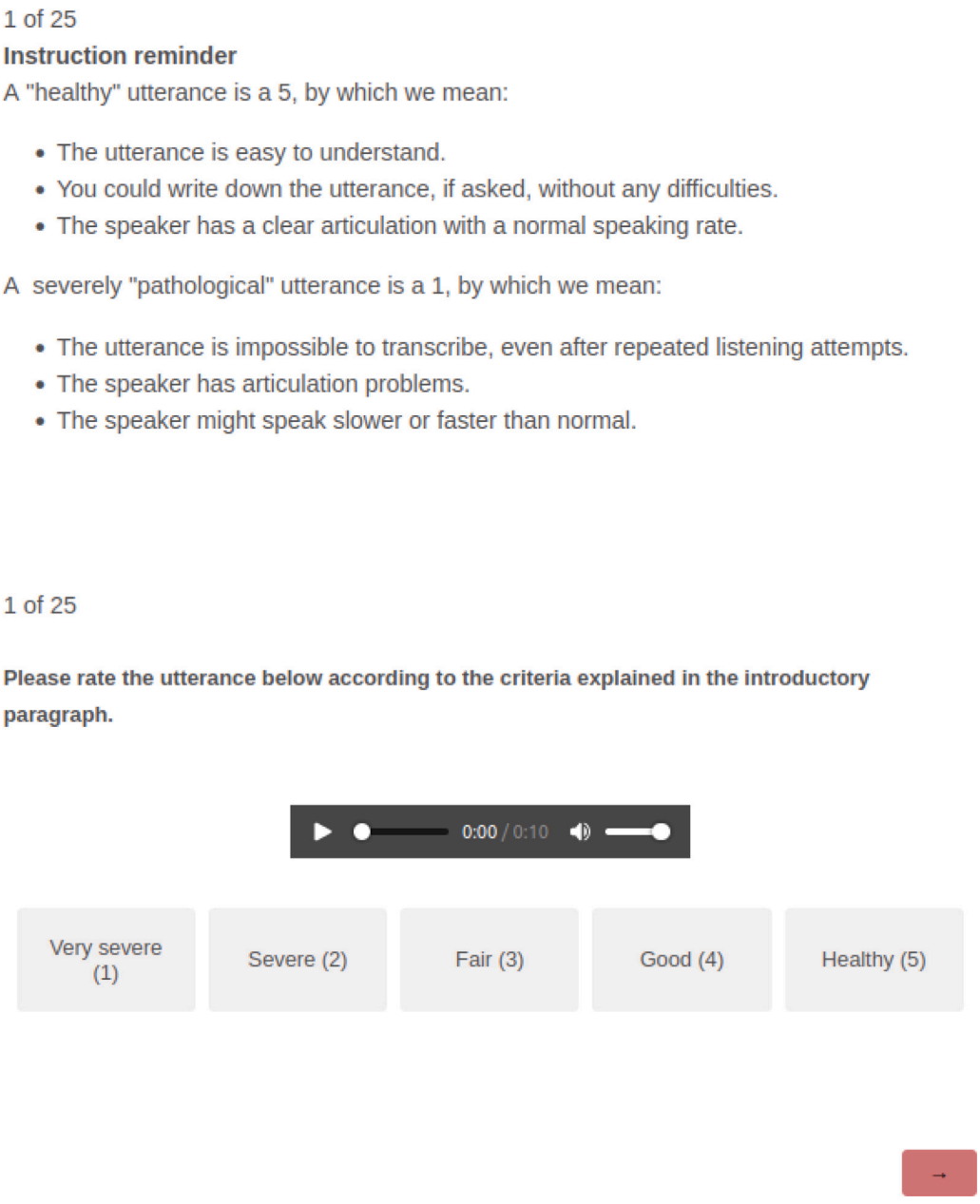

Fig. 4. Screenshot from the graphical user interface of the questionnaire.

voice quality. This would actually be in line with previous findings from our lab that showed that although ASR-based measures are often said to measure the intelligibility, we observed that extralinguistic cues can also influence ASR performance (Feng et al., 2021). Therefore, this study would like to also communicate that there is an interesting direction in testing ASR tools for estimating voice quality or naturalness tools for estimating severity of the speech disorder.

Having been influenced by the idea of testing naturalness measures for evaluating severity, we tested three methods that are traditionally used for synthetic speech evaluation: the GV detector/regressor, the MS detector/regressor, and the MCD. In our results, we found that the GV and the MS feature based models both performed very well in comparison to the other features that we have tested: On the reference-free evaluation, the MS-regressor had the highest correlation with the listener ratings, and the GV-regressor had the second highest correlation. On the reference-based evaluation, the MS-regressor was also the best performing. These results indicate that speech synthesis evaluation approaches are working well for the evaluation of speech severity.

Based on this, we think naturalness measures could be used for evaluating speech severity. To further improve speech severity evaluation measures, we believe that a better understanding of the concepts of severity and naturalness would be needed. In our previous studies we found that this distinction is often not clear for naive listeners (Halpern et al., 2021; Illa et al., 2021). It would be interesting to compare how the ratings of listeners would differ when rating the same set of stimuli both for severity and naturalness, i.e., would there be statistically significant difference in the ratings? If not, that means that the psychometric questionnaires measuring these values should be rethought, as naturalness and severity should be rated differently.

Our second research question concerned whether there are other approaches available that require less labelled training data while giving similar performance on the speech evaluation task. The naive listener experiment in Section 3.5 showed that the severity ratings of naive listeners have a very high correlation with the expert listener's severity ratings. These results imply that it is possible to reliably, and cost-efficiently scale up the annotation of oral cancer speech for

the prototyping of data-driven automatic objective speech severity evaluation approaches. An obstacle towards using crowdsourcing in clinical scenarios is that informed consent will have to be obtained from patients for these rating studies. Currently no speech data needs to be shared with strangers other than the hospital staff, who rate these utterances internally. Some patients might feel uncomfortable sharing their speech with strangers participating in these studies.

The high correlation of expert and naive listener scores are partially contrary to the findings of Carvalho et al. (2021) who found that expert listeners were significantly better at transcription of dysarthric utterances. The differences in the findings could be explained by the difference in the transcription task. The study of Carvalho et al. (2021) used a transcription task to quantify intelligibility while we chosen a more impressionistic severity measure where the perceived intelligibility was just one of the factors to be rated. It could be that our raters overestimated their ability to transcribe oral cancer speech as they have not been explicitly asked to transcribe the utterances.

We also found that using binary labels indicating the presence or absence of oral cancer speech led to a reduced labelling effort but also nearly always resulted in a lower correlation with the human ratings than using the full 5-point scale ratings. For severity ratings, we thus advice to use a graded scale rather than binary labels.

Apart from the possible clinical application of our method outlined above, we think that our results are potentially good enough for use in smartphone applications: (1) The ASR and the LASSO models presented here have relatively low computational complexity compared to fully deep learning based methods such as Quintas et al. (2020), which is important due to the low memory requirements of smartphone devices. (2) In a smartphone use case, various noises and unexpected (conversational, spontaneous) speech modalities can be present. As our approaches have been tested with these scenarios, we are confident that performance will not deteriorate significantly in these conditions. Still, a more controlled test would be imminent, where similar speech recordings should be tested under different, real life, controlled noise conditions. For these smartphone scenarios, we suggest using the modulation spectrum regressor, if no reference transcription is available, and the DNN AM Retraining+ngram ASR model when a reference transcription is available. We assumed that all the features were already extracted in the present study. However, these would have to be extracted on-device in a smartphone. Taking the extraction into consideration, the MS and the GV features would be faster on a smartphone than the x-vector or the d-vector. Extraction of the latter two features would require an additional pass through the neural network on top of the spectrogram calculation.

Finally, the ASR-based methods in this paper could be potentially further improved by considering a spontaneous dataset in the pretraining stage. During the time of experimentation, we were only aware of the Switchboard dataset, however this had a lower sampling frequency than the WSJ, which we have deemed more important than the matched modality of the speech. Pretraining with, e.g., the HUB4 dataset, which consists of prepared and spontaneous journalistic speech, could be a lucrative direction for improvement.

7. Conclusion

In this paper, we aimed to find the best method for the automatic evaluation of the severity of oral cancer speech. To do that, we collected a publicly available spontaneous oral cancer speech corpus. We compared two sets of reference-based methods and one set of reference-free methods.

Our extensive experiments showed:

(1) an ASR model was found to have the highest correlation with the human ratings, when we have access to a transcription (reference-based): the DNN AM retraining model, an ASR model which uses oral cancer data during training and an n-gram based language model. (2) When we do not have access to a transcription a LASSO regression model was found to be the best using modulation spectrum features. (3) In an effort to reduce labelling effort, we found that naive listeners' ratings, e.g., obtained through crowd-sourcing, can be used instead of those of an expert listeners as their ratings were highly similar. Therefore, we encourage the usage of naive listener scores for speech severity labelling to reduce data collection costs, and therefore prototype automatic speech severity evaluation systems more efficiently. (4) We found that the use of binary labels led to lower correlations of the automatic methods than using the severity scores.

CRediT authorship contribution statement

Bence Mark Halpern: Conceptualization, Data curation, Investigation, Formal analysis, Methodology. **Siyuan Feng:** Methodology, Formal analysis, Investigation. **Rob van Son:** Writing – review & editing, Funding acquisition, Supervision. **Michiel van den Brekel:** Writing – review & editing, Funding acquisition, Supervision. **Odette Scharenborg:** Writing – review & editing, Supervision, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Dataset link is given in the article

Acknowledgements

We would like to thank Noa Hannah for helping out with the SLP ratings. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No 766287. The Department of Head and Neck Oncology and surgery of the Netherlands Cancer Institute receives a research grant from Atos Medical (Hörby, Sweden), which contributes to the existing infrastructure for quality of life research.

Appendix

A.1. Questionnaire instructions used in the rating study

Dear Sir/Madam

In this task, you will listen to recordings containing one utterance each. The speakers in these recordings all had oral cancer. They are going to talk about their experience with oral cancer.

We ask you to rate each utterance according to the severity of the oral cancer impact on the spoken utterance, i.e., we ask you to rate how "pathological" the utterance is.

A "healthy" utterance is a 5, by which we mean:

- *The utterance is easy to understand.*
- *You could write down the utterance, if asked, without any difficulties.*
- *The speaker has a clear articulation with a normal speaking rate.*

A severely "pathological" utterance is a 1, by which we mean:

- *The utterance is impossible to transcribe, even after repeated listening attempts.*
- *The speaker has articulation problems.*

- *The speaker might speak slower or faster than normal.*

In total, you are going to listen to 25[8] *utterances. Each utterance is about 10 s long. You can listen to the utterances as many times as you want. Your task is to rate the severity of the pathology of the utterances.*

On the next page, we are going to present you with a "healthy" and a "pathological" utterance (see Fig. 4).

References

American Speech Language Hearing Association, (2023). Voice disorders. URL: https://www.asha.org/practice-portal/clinical-topics/voice-disorders/#collapse_5.

Balaguer, M., Boisguerin, A., Galtier, A., Gaillard, N., Puech, M., Woisard, V., 2019. Factors influencing intelligibility and severity of chronic speech disorders of patients treated for oral or oropharyngeal cancer. Eur. Arch. Oto-Rhino-Laryngol. 276 (6), 1767–1774.

Bin, L., Kelley, M.C., Aalto, D., Tucker, B.V., 2019. Automatic speech intelligibility scoring of head and neck cancer patients with deep neural networks. In: International Congress of Phonetic Sciences (ICPHs' 19), Melbourne, Australia. pp. 3016–3020.

Bredin, H., Yin, R., Coria, J.M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., Gill, M.-P., 2020. Pyannote. audio: neural building blocks for speaker diarization. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 7124–7128.

Carvalho, J., Cardoso, R., Guimarães, I., Ferreira, J.J., 2021. Speech intelligibility of Parkinson's disease patients evaluated by different groups of healthcare professionals and naïve listeners. Logop. Phoniatr. Vocology 46 (3), 141–147. http://dx.doi.org/10.1080/14015439.2020.1785546, PMID: 32633172.

Coria, J.M., Bredin, H., Ghannay, S., Rosset, S., 2020. A comparison of metric learning loss functions for end-to-end speaker verification. In: Espinosa-Anke, L., Martín-Vide, C., Spasić, I. (Eds.), Statistical Language and Speech Processing. Springer International Publishing, pp. 137–148.

Dagenais, P.A., Brown, G.R., Moore, R.E., 2006. Speech rate effects upon intelligibility and acceptability of dysarthric speech. Clin. Linguist. Phon. 20 (2–3), 141–148.

Duffy, J.R., 2005. Motor Speech Disorders: Substrates, Differential Diagnosis and Management, Vol. 1. p. 96.

Feng, S., Kudina, O., Halpern, B.M., Scharenborg, O., 2021. Quantifying bias in automatic speech recognition. arXiv preprint arXiv:2103.15122.

Foundation, O.C., 2019. Oral cancer facts. URL: https://oralcancerfoundation.org/facts/.

Gales, M.J., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. 12 (2), 75–98.

Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., Khudanpur, S., 2014. A pitch extraction algorithm tuned for automatic speech recognition. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 2494–2498.

Halpern, B.M., Feng, S., van Son, R., van den Brekel, M., Scharenborg, O., 2022. Low-resource automatic speech recognition and error analyses of oral cancer speech. Speech Communication 141, 14–27.

Halpern, B.M., Fritsch, J., Hermann, E., van Son, R., Scharenborg, O., Magimai-Doss, M., 2021. An objective evaluation framework for pathological speech synthesis. In: Speech Communication; 14th ITG Conference. VDE, pp. 1–5.

Halpern, B.M., van Son, R., van den Brekel, M., Scharenborg, O., 2020. Detecting and analysing spontaneous oral cancer speech in the wild. In: Proc. Interspeech 2020. pp. 4826–4830. http://dx.doi.org/10.21437/Interspeech.2020-1598.

Hirano, M., 1981. GRBAS" scale for evaluating the hoarse voice & frequency range of phonation. Clin. Exam. Voice 5, 83–84.

Huang, W.-C., Halpern, B.M., Violeta, L.P., Scharenborg, O., Toda, T., 2022. Towards identity preserving normal to dysarthric voice conversion. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 6672–6676.

Illa, M., Halpern, B.M., van Son, R., Moro-Velázquez, L., Scharenborg, O., 2021. Pathological voice adaptation with autoencoder-based voice conversion. In: 11th ISCA Speech Synthesis Workshop. ISCA, pp. 19–24.

Jacobi, I., van der Molen, L., Huiskens, H., Van Rossum, M.A., Hilgers, F.J., 2010. Voice and speech outcomes of chemoradiation for advanced head and neck cancer: a systematic review. Eur. Arch. Oto-Rhino-Laryngol. 267 (10), 1495–1505.

Janbakhshi, P., Kodrasi, I., Bourlard, H., 2019. Pathological speech intelligibility assessment based on the short-time objective intelligibility measure. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 6405–6409.

Kim, H., 2015. Familiarization effects on consonant intelligibility in dysarthric speech. Folia Phoniatr. Logop. 67 (5), 245–252.

Kim, H., Nanney, S., 2014. Familiarization effects on word intelligibility in dysarthric speech. Folia Phoniatr. Logop. 66 (6), 258–264.

Kominek, J., Black, A.W., 2004. The CMU Arctic speech databases. In: Proc. 5th ISCA Workshop on Speech Synthesis (SSW 5). pp. 223–224.

Kubichek, R., 1993. Mel-cepstral distance measure for objective speech quality assessment. In: Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, Vol. 1. IEEE, pp. 125–128.

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. In: Interspeech. ISCA, pp. 2943–2947.

Laaridh, I., Kheder, W.B., Fredouille, C., Meunier, C., 2017. Automatic prediction of speech evaluation metrics for dysarthric speech. In: Proc. Interspeech 2017. pp. 1834–1838. http://dx.doi.org/10.21437/Interspeech.2017-1363.

Lansford, K.L., Borrie, S.A., 2017. Use of crowdsourcing platforms to examine listener perception of disordered speech. J. Acoust. Soc. Am. 141 (5), 3911.

Lansford, K.L., Borrie, S.A., Bystricky, L., 2016. Use of crowdsourcing to assess the ecological validity of perceptual-training paradigms in dysarthria. Am. J. Speech-Lang. Pathol. 25 (2), 233–239.

Maier, A., Haderlein, T., Stelzle, F., Nöth, E., Nkenke, E., Rosanowski, F., Schützenberger, A., Schuster, M., 2009. Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer. EURASIP J. Audio Speech Music Process. 2010, 1–7.

Martínez, D., Green, P.D., Christensen, H., 2013. Dysarthria intelligibility assessment in a factor analysis total variability space. In: Proc. Interspeech 2013. pp. 2133–2137. http://dx.doi.org/10.21437/Interspeech.2013-505.

Mashimo, M., Toda, T., Shikano, K., Campbell, N., 2001. Evaluation of cross-language voice conversion based on GMM and straight. In: Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001). pp. 361–364.

Master, S., De Biase, N., Pedrosa, V., Chiari, B.M., 2006. The long-term average spectrum in research and in the clinical practice of speech therapists. Pro-Fono : Rev. Atualizacao Cient..

McFee, B., Lostanlen, V., Metsai, A., McVicar, M., Balke, S., Thomé, C., Raffel, C., Zalkow, F., Malek, A., Dana, Lee, K., Nieto, O., Mason, J., Ellis, D., Battenberg, E., Seyfarth, S., Yamamoto, R., Choi, K., viktorandreevichmorozov, Moore, J., Bittner, R., Hidaka, S., Wei, Z., nullmightybofo, Hereñú, D., Stöter, F.-R., Friesch, P., Weiss, A., Vollrath, M., Kim, T., 2020. Librosa/librosa: 0.8.0. http://dx.doi.org/10.5281/zenodo.3955228.

Meyer, T.K., Kuhn, J.C., Campbell, B.H., Marbella, A.M., Myers, K.B., Layde, P.M., 2004. Speech intelligibility and quality of life in head and neck cancer survivors. Laryngoscope 114 (11), 1977–1981.

Nagrani, A., Chung, J.S., Huh, J., Brown, A., Coto, E., Xie, W., McLaren, M., Reynolds, D.A., Zisserman, A., 2020. Voxsrc 2020: The second voxceleb speaker recognition challenge. arXiv preprint arXiv:2012.06867.

Oates, J., 2009. Auditory-perceptual evaluation of disordered voice quality. Folia Phoniatr. Logop. 61 (1), 49–56.

O'Sullivan, B., Shah, J., 2003. New TNM staging criteria for head and neck tumors. In: Seminars in Surgical Oncology. Wiley Online Library, pp. 30–42.

Paul, D.B., Baker, J.M., 1992. The design for the wall street journal-based CSR corpus. In: Proceedings of the Workshop on Speech and Natural Language. Association for Computational Linguistics, pp. 357–362.

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. In: Proc. Interspeech 2020. pp. 4976–4980. http://dx.doi.org/10.21437/Interspeech.2020-1431.

Ravanelli, M., Bengio, Y., 2018. Speaker recognition from raw waveform with sincnet. In: 2018 IEEE Spoken Language Technology Workshop. SLT, IEEE, pp. 1021–1028.

Revis, J., Giovanni, A., Wuyts, F., Triglia, J.-M., 1999. Comparison of different voice samples for perceptual analysis. Folia Phoniatr. Logop. 51 (3), 108–116.

Rudzicz, F., Namasivayam, A.K., Wolff, T., 2012. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. Lang. Resour. Eval. 46 (4), 523–541.

Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al., 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 4779–4783.

Shield, K.D., Ferlay, J., Jemal, A., Sankaranarayanan, R., Chaturvedi, A.K., Bray, F., Soerjomataram, I., 2017. The global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2012. CA: Cancer J. Clin. 67 (1), 51–64.

Smith, L.K., Goberman, A.M., 2014. Long-time average spectrum in individuals with Parkinson disease. NeuroRehabilitation http://dx.doi.org/10.3233/NRE-141102.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S., 2018. X-vectors: Robust dnn embeddings for speaker recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 5329–5333.

Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 4214–4217.

Takamichi, S., Toda, T., Neubig, G., Sakti, S., Nakamura, S., 2014. A postfilter to modify the modulation spectrum in HMM-based speech synthesis. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 290–294.

[8] This was 24 in one of the questionnaires.

Tanner, K., Roy, N., Ash, A., Buder, E.H., 2005. Spectral moments of the long-term average spectrum: Sensitive indices of voice change after therapy? J. Voice http://dx.doi.org/10.1016/j.jvoice.2004.02.005.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.

Toda, T., Tokuda, K., 2007. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. IEICE Trans. Inf. Syst. 90 (5), 816–824.

Tripathi, A., Bhosale, S., Kopparapu, S.K., 2020. A novel approach for intelligibility assessment in dysarthric subjects. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 6779–6783.

Union, I.T., 1996. Methods for subjective determination of transmission quality.

Vincent, E., Gribonval, R., Févotte, C., 2006. Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. 14 (4), 1462–1469.

Wan, L., Wang, Q., Papir, A., Moreno, I.L., 2018. Generalized end-to-end loss for speaker verification. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 4879–4883.

Ward, E.C., van As-Brooks, C.J., 2014. Head and Neck Cancer: Treatment, Rehabilitation, and Outcomes. Chapter 5: Speech and Swallowing Following Oral, Oropharyngeal, and Nasopharyngeal Cancer. Plural Publishing.

Windrich, M., Maier, A., Kohler, R., Nöth, E., Nkenke, E., Eysholdt, U., Schuster, M., 2008. Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma. Folia Phoniatr. Logop. 60 (3), 151–156.

Woisard, V., Balaguer, M., Fredouille, C., Farinas, J., Ghio, A., Lalain, M., Puech, M., Astesano, C., Pinquier, J., Lepage, B., 2022. 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.

Wolfe, V., Cornell, R., Fitch, J., 1995. Sentence/vowel correlation in the evaluation of dysphonia. J. Voice 9 (3), 297–303.

Yamamoto, R., Yamada, Y., hyama5, Aria-K-Alethia, Huang, R., Hiroshiba, Regan, J., Roszkowski, M., Shirani, T., 2021. R9y9/nnmnkwii: v0.1.0 release. http://dx.doi.org/10.5281/zenodo.5178769.

Zhou, X., Garcia-Romero, D., Mesgarani, N., Stone, M., Espy-Wilson, C., Shamma, S., 2012. Automatic intelligibility assessment of pathologic speech in head and neck cancer based on auditory-inspired spectro-temporal modulations. In: Proc. Interspeech 2012. pp. 542–545. http://dx.doi.org/10.21437/Interspeech.2012-105.

Zraick, R.I., Kempster, G.B., Connor, N.P., Thibeault, S., Klaben, B.K., Bursac, Z., Thrush, C.R., Glaze, L.E., 2011. Establishing validity of the consensus auditory-perceptual evaluation of voice (CAPE-V). Am. J. Speech-Lang. Pathol..